ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Ernst & Young

Expected Shortfall Backtesting

Econometrics and Management Science

Quantitative Finance

Master's Thesis - FEM21031-18

Florentijn Gelling (400295)

Supervisors: Karolina Scholtus (EUR)

Niels van der Kleij (EY)

Second assessor: Prof. Dr. Chen Zhou (EUR, DNB)

25th July 2019

Abstract

This paper is on the change in risk measure from Value at Risk (VaR) to Expected Shortfall (ES) as stipulated by the Basel Committee of Banking Supervision, and in particular, the backtesting methods that have been developed for this new risk measure ES. The contribution of this research to the literature is an analysis of the performance of the most relevant contemporary backtests in several different modelling scenarios and a more simple scenario where their performances are compared to the currently used VaR setting. We analyse the (size-adjusted) power of six different methods in a variety of statistical modelling scenarios, and comment on computational time and implementational complexity of these methods. This is done for the consideration of widespread use throughout the financial system, and in particular for regulatory purposes. We observe that both the methods by Graham and Pál (2014) (GP) and by Moldenhauer and Pitera (2018) (MP) outperform the other methods in terms of size-adjusted power throughout the analyses. In terms of computational time, the GP method beats out the calibration time necessary for the MP method, though the MP method performs (daily) evaluations much faster, once the initial calibration is done. Furthermore, the MP method is slightly more consistent than the GP method, and it is much more easily implemented due to its relatively low mathematical complexity. Therefore, we recommend its use as the primary methodology for backtesting ES in matters related to risk management in the financial sector. Lastly, the MP method seems to be slightly more sensitive to misspecifications than the currently used VaR backtest is in the VaR framework. This suggests that, if the MP method is to be applied, the quantity and size of fees for inadequate risk management will increase for financial services organisations, if their current models remain in use.

Keywords: *Value at Risk, Expected Shortfall, Backtesting, Tail risk management*

Acknowledgements

This research has been performed with the help and guidance of Karolina Scholtus from the Erasmus School of Economics, Prof. Dr. Chen Zhou from the Erasmus School of Economics and the Risk Management team of EY Amsterdam, in particular Niels van der Kleij and Rens Borsje. All four of them have been instrumental in the shaping of the overall direction of the research and the smoothing of the rough edges, as well as being a great source of knowledge on all topics concerning the contents of this thesis.

Contents

1 Introduction

1.1 Background

Expected Shortfall (ES) is a risk measure that captures tail risk better than the Value at Risk (VaR) measure that has been used for decennia. Despite the fact that regulatory authorities have been reluctant to phase in this new risk measure in favour of the old one, possibly supported by lobbying financial institutions, this is not to last. At the moment, discussion is still being held on the specifics, but current estimates place this change around early 2022.

The Basel Committee on Banking Supervision stipulates that financial institutions will be required to report daily ES estimates, using 97.5^{th} percentile, one-tailed confidence levels (BCBS, 2016). Before this is fully implemented, sound ways of backtesting these estimates must be developed. This is, however, not as easy as one might think, coming from a VaR setting[1].

Before discussing the VaR framework, we must understand what sort of data these tail risk measurements are based on. For the scope of this research, we will use profits and losses (P&Ls) to indicate the value of an investment. These P&L values are first differences of logarithms of spot prices of stocks and derivatives, or a combination of several spot prices gathered in a portfolio. Since VaR values are not additive, and neither are those of ES, portfolio risk measures are quite complicated to construct, especially since lognormal returns are considered, which do not add up to any known distribution. The current preferred way of dealing with this is by use of the Variance-Covariance method, which is discussed in greater detail in Section 3.2.4.

We define the VaR risk measure as follows:

$$\mathrm{VaR}_\nu(X) = \inf\{x | F_X(x) \geq \nu\}. \tag{1}$$

Some researchers define their VaR measure as the negative of Equation 1, such that the 'daily loss' is artificially represented as a positive value; we will not. This ties into the choice for $\nu = 0.01$ for VaR and $\nu = 0.025$ for ES, as the aforementioned researchers generally define ν to be 0.99 for VaR and 0.975 for ES. We view the distribution as a whole profit and loss statement, where the losses are on the left side of the curve (and are in fact negative numbers), of which we take the 2.5^{th} percentile. As per illustration, we refer to (Dimitriadis & Bayer, 2017), who utilise the same notation as we do.

For this risk measure multiple backtesting methods are proposed, they fall into one of the following three categories (Holton, 2014):

- **Coverage tests** - assess whether the frequency of exceedances is consistent with the loss quantile that the VaR level intended to reflect;
- **Distribution tests** - apply goodness-of-fit tests to overall loss distribution estimates;
- **Independence tests** - assess independence of results from one period (usually daily) to the next.

A good VaR prediction model should pass all these testing criteria, since these are complementary and not interchangeable. The most straightforward of these types is the Coverage test, which is why we

[1] Which uses a 99^{th} percentile, one-tailed confidence level

will focus on this type of test for now. This backtest is also relatively easy to perform using known test statistics (Chi-square for instance). The reason this works is that the VaR measure (in this test setup) only cares about whether or not a P&L observation of a given day exceeds the threshold pertaining to a specific VaR-percentile. Therefore, it can easily be shown that the total amount of exceedances is Bernoulli distributed, assuming exceedances are independent (which requires confirmation from an Independence test) and that the only necessary parameters for the probability density function (pdf) are the sample size and the probability of an exceedance occurring. This gives an easy way to perform a test on whether the predicted VaR threshold was reasonable, given the observations, since we simply need to test the null hypothesis that the true exceedance proportion equals that which the VaR-percentile suggests.

One of the greatest strengths of VaR is its simplicity in backtesting through its Bernoulli distribution (although Independence and Distribution tests are also necessary from a theoretical perspective). However, one of the problems with the VaR measure is that it gives us no insight whatsoever in our tail risk beyond the threshold value. It tells us the magnitude of the loss that will not be exceeded 1% of the time, but it gives no details on how extreme the loss can get in those 1% outlier cases. Furthermore, if the VaR backtest fails, the measure contains no information or suggestion on what the correct VaR level truly is. As opposed to giving just a threshold for risk, the ES risk measure takes into account what happens when the threshold is passed. It is the expected loss, given that this loss exceeds a certain threshold:

$$\mathrm{ES}_\nu(X) = \frac{1}{\nu}\int_0^\nu VaR_\mu(X)d\mu. \tag{2}$$

Through this fact, however, it cannot simply be backtested the way VaR is. Even more problematic, though, is that ES is only concerned with P&L observations in this VaR-percentile, which means that only the exceedances are relevant information for the ES measure, and all other observations above the threshold are meaningless. Since Basel stipulates the use of the 97.5^{th} percentile, backtesting over a period of 1 year (252 trading days) gives us, by construction, an expected amount of useful data of about 6 observations. This small-sample problem is an issue that renders standard significance tests useless, since testing power is in general heavily correlated with sample sizes.

As a practical illustration, we see that even a simple test for $\mu_0 = 0$ (population mean under null hypothesis) requires an observation average that lies very far from zero in order to be rejected with a confidence level of $1 - \alpha = 95\%$, when using a sample size of $n = 6$, shown here:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}, \quad \text{which gives:} \quad \overline{x}/\sigma \geq \frac{1.96}{\sqrt{6}} \approx 0.80, \tag{3}$$

which means that the average of the 6 observations needs to be 0.8 standard deviations off $\mu_0 = 0$ to reject μ_0, but this is only true if the data is normally distributed and the variance of the population is known. If the variance is unknown, we must use the sample variance S and the t-statistic with 5 degrees of freedom (dof, denoted as θ), which gives a coefficient of 2.571. This results in an $\overline{x}/\sigma$ value of 1.04, meaning that the observed average must be more than one standard deviation distance from μ_0 in order

to reject it, showing that the average of the data must be at least 25% more extreme in order to reject the null hypothesis, if we do not know the underlying distribution, in this particular setup. To make matters worse, P&L data has been repeatedly shown to be fat-tailed, so normal assumptions are rather unrealistic for ES prediction and backtesting (Graham & Pál, 2014), (McNeil, Frey, Embrechts et al., 2005). Note that this is also an issue for the VaR risk measure. So, if we also cannot rely on our data to follow a normal distribution, the above test is only valid in approximation using the central limit theorem, with the extra necessity of $n \geq 30$. Analysing this situation shows us that standard methods like the one utilised above are rather unfeasible in this research, and they require very extreme tail loss observations for rejection of ES estimates to be confidently asserted.

When backtesting ES note that some methods assume the VaR level to be correct, and thus are insensitive to large amounts of exceedances, while other methods implicitly test for the VaR level in their procedure for testing ES. The ES measure has the same amount of violations as the VaR measure, by construction, but ES takes the size of the loss into account too. Not all of the ES backtests, however, actually also take the quantity of violations into account. Naturally, one must remember to evaluate the VaR level separately, when relying on this type of ES backtests. Unless otherwise noted, we will generally assume the VaR threshold to be correctly specified.

Graham and Pál (2014) give a very clear and concise summary of p-values, hypothesis testing and rejection regions, which will be used rather extensively in our analysis. This, however, is assumed common knowledge within the field of econometrics and quantitative finance and will not be discussed in more detail.

1.2 General notes on ES backtesting

1.2.1 Simulation and empirical analysis

When researching the validity of any backtesting procedure using historical simulation, one cannot discriminate between the two possible outcome scenarios, both when the backtest accepts the predictions and when the backtest rejects the predictions. If the backtest accepts the prediction, this could be because the prediction model was correct and the backtest is correct (or well enough calibrated to this specific scenario), or it could be because the prediction model specification was incorrect, but the backtest is also inaccurate/incorrect/misspecified/miscalibrated or simply put, flawed. Thus, either both were right or two wrongs might make a right in terms of the significance test outcome. On the other hand, if the backtest rejects the prediction this could be because the prediction model was incorrect and the backtest is correct (for this specific scenario) such that it correctly rejects. The alternative is that the prediction model was correctly specified (or close enough), but the backtest is flawed. This results in one wrong and one right that together make a wrong significance test outcome; without additional information, there is no way to tell which is which.

In order to get a 'clean' evaluation of the validity of a backtesting method, one must thus know whether the prediction model was specified correctly or not. Only then can we know whether the

backtest performs desirably or not. This is relatively easily done using simulated data, since this allows a researcher to control, and thus know both the correct model specification and the true ES through analytical calculation or MC methods.

We will analyse the difference in sensitivity to incorrect specifications between a common and accepted VaR backtest and our ES backtests of interest. This can be instrumental in recognising, understanding and dealing with the problems that come along with this structural change in the financial reporting through the move of the Fundamental Review of the Trading Book (FRTB) from VaR to ES. We can do this relatively easily since the VaR 99^{th} and ES 97.5^{th} percentile values are (almost) identical for a normal distribution (Moldenhauer & Pitera, 2018). Thus, we can analyse whether the VaR backtests start rejecting first, or the ES backtests, when model misspecifications get gradually more extreme.

1.2.2 Regulatory versus Corporate

We note a specific distinction between the regulator's perspective and that of financial institutions, e.g. banks, on the backtesting of tail risk.

First off, from a bank's perspective we would like to monitor tail risk as accurately as possible, such that we know exactly how much liquidity we need to hold in reserve to cover our risk, without being too conservative with our assets and missing out on potential extra revenue. From a regulatory perspective we wish to monitor tail risk as accurately as possible, to keep the financial system in check, such that no large scale systemic risks are taken that endanger the economy as a whole. On the other hand, we must not hinder or obstruct the financial system unnecessarily, such that it runs smoothly and does not suffer inefficiencies. This is a balance on a knife's edge due to the interactions between the two involved parties. From a bank's perspective the main objective is profit, whereas from a regulatory perspective the main objective is risk mitigation[2]. We see that both have the same immediate goal of tail risk measurement, and in an ideal world the preferences of the two should line up perfectly. However, given that measuring risk is not an exact science (through its dependency on distributional assumptions) and the fact that the involved parties do not have the same underlying reasons for their interest in tail risk, the financial institutions will inevitably move towards the more risky side of the spectrum of risk assessment, while regulators will move to the more conservative side.

Secondly, financial institutions develop highly specific and complex models for their tail risk assessment. We can make our backtesting method highly model-specific by attuning it to one such underlying prediction model, such that it calculates exactly the probability of the underlying model being correct, given the observed data. On the other hand, regulatory authorities are only really interested in the end result, the discrepancy between the predicted and the observed tail loss. Thus, the regulator would ideally employ a backtesting methodology that can be used for all tail risk assessments in exactly the same way. This method would simply take ES predictions and observations of a given year as input, and

[2]An argument can be made that a financial institution would rather have a 'healthy' risk assessment than maximise profit at all costs, but the most important detail is that the financial institutions' preferences will always be on the risky side *compared to* the preferences of the regulator.

produce a p-value as a result, regardless of the detail and complexity of the prediction model and error distribution assumptions concerned.

In this research, preference is given to the regulatory perspective, for its much wider applicability and usefulness compared to a backtesting method that is highly tailored to the tail risk measurement of some singular financial institution. With that said, most of the methods still require the specification of an assumed underlying pdf (empirical, normal, t, etc.). To deal with this, the methods are programmed such that the distributional specification can be changed on the fly.

1.2.3 Quantifying the performance of a backtest

We discuss and define what a 'good' backtest consist of in this section. Size and power properties are evaluated in virtually every piece of literature on the subject[3]. To these two measures we add the issue of ease and appropriateness of widespread implementation. We list the four as follows:

- Power: sensitivity to incorrect specification (for low type 2 error);
- Size: acceptance rate for correct (or close enough) specification (for low type 1 error);
- Ease of implementation/complexity issue (for wide use by regulators and the financial system);
- Applicability in terms of what setting it works desirably/correctly for (for wide use by regulators and the financial system).

Furthermore, size-adjusted power levels (Lloyd, 2005) will also be considered for the comparison of methods in terms of power when their sizes are unequal.

1.3 Research question

This research is on the topic of methods for backtesting ES. The research concerns the following methods: Del Brio, Mora-Valencia and Perote (2017)'s t-test, M. Righi and Ceretta (2013)'s truncated distribution dispersion, Graham and Pál (2014)'s saddle-point method, Löser, Wied and Ziggel (2018)'s Irwin-Hall transformation, Moldenhauer and Pitera (2018)'s secured position and Bayer and Dimitriadis (2018)'s regression-based backtest.

The research focuses on the following question: 'How do the methods mentioned above compare for the purpose of backtesting Expected Shortfall (in terms of size, power, complexity and applicability) when used by regulators as well as the financial system as a whole?'

This can be analysed in a simulated environment (Wimmerstedt, 2015), (Löser et al., 2018) and (Bayer & Dimitriadis, 2018). However, if we are to get an indication of what impact the backtesting of ES over that of VaR has, the performance of these backtests must be compared to the original VaR backtest, in a common scenario that the latter is used in in practice. Apart from analysing which backtests perform best in terms of power in a statistically complex environment, we hereby also research

[3]We follow Bayer and Dimitriadis (2018), who define the size of a test as the rejection frequency of the test under the null hypothesis, which should equal the nominal signicance level, and the power of a test as the rejection frequency of forecasts stemming from some misspecied model, which is optimally as close to one as possible.

how these methods will change the field of tail risk management, when the switch to the ES risk measure is fully realised.

The simulation scenarios consist of a data generating process (dgp) under a range of parametric specifications (e.g. t-distributions with differing degrees of freedom) and predictions under a range of hypotheses following the same parametric specifications. We deliberately both correctly and incorrectly match these dgp's with the predictions under certain hypotheses, in order to create several scenarios of which we know a priori whether the predictions are theoretically correct or incorrect for the applied dgp. Following this we apply the backtesting methods and observe whether these indicate rejection of the prediction method used or not. In case that the method was matched correctly, we expect a rejection rate of $\alpha = 5\%$. Alternatively, if the dgp and the prediction method were mismatched, we expect a higher rejection rate, up to a limit of 100% rejection, depending on the severity of the mismatch[4]. From this we can conclude which backtest is the most powerful across the defined scenarios.

Applying the above, we have a methodology at our disposal for researching the validity of ES backtests in several parametric settings (following any arbitrary distribution). Furthermore, we can compare these ES backtests with the currently employed method of analysing tail risk (which is VaR backtesting), with the goal of researching which of the considered backtests is the preferred one for widespread implementation.

2 Literature

When it comes to available literature, quite a lot of material on backtesting ES has been produced since the turn of the millennium. After all, the Basel committee admitted that the difficulty of replacing VaR by ES, which will be required by Basel soon, lies within the backtesting of prediction models (Dalne, 2017). Most of the proposed methods, however, have some caveats. Many of them depend heavily on underlying distributional assumptions, which are hardly appropriate in a general application given the documented non-normality of stock returns, or for financial excess returns in general for that matter (Sheikh & Qiao, 2009). Section 2.1 focuses on what backtesting methods have been proposed, giving a short description of the inner workings of those that are relevant to us and highlighting the pros and cons in terms of applicability and assumptions used.

Given that we are especially interested in finding ways of backtesting ES predictions regardless of what model or method was used in constructing these predictions, we will have to research how this lack of validity of the underlying assumptions affects the power of such methods, or we must employ methods that do not rely on these assumptions at all. Thus, in section 2.2, we will discuss how the power of a backtest can be measured, according to previous literature.

[4]We illustrate this with the following example: when the dgp follows a t-distribution with 3 dof, we expect a much larger rejection rate in the scenario where the null hypothesis is of a t-distribution with 30 dof than a scenario where the null hypothesis is of a t-distribution with 5 dof, because the second hypothesis is much closer to the underlying data than the first.

2.1 Existing backtests of ES

The first methods that were developed for ES backtesting were published in the early 2000's. McNeil and Frey (2000) proposed a residual approach, taking the sum of the differences between ES predictions and actual tail loss observations, scaled by volatility; and employing Extreme Value Theory (EVT). After this, Kerkhof and Melenberg (2004) came up with the so-called functional delta method. Both of these methods unfortunately rely on asymptotic statistics, which are inaccurate when dealing with small sample sizes (Dalne, 2017). Other earlier methods include Berkowitz (2001)'s censored Gaussian approach and Bradley and Taqqu (2003).

These earlier methods were not equipped to deal with small sample sizes, nor were they capable of accurate predictions under non-normality. Wimmerstedt (2015) reviews the methods of Wong (2008), M. Righi and Ceretta (2013), Acerbi and Szekely (2014) and Emmer, Kratz and Tasche (2015), which we will discuss next.

We found an issue with Wimmerstedt's execution of the analysis on accepting true predictions for the method by Emmer et al. (2015). Because of this error, Wimmerstedt chooses in favour of the other methods over Emmer's. The problem itself lies in the rejection region of a discrete distribution. In the case of continuous distributions, rejecting an observation if it is above the 95% cumulative density threshold $[p(X \leq x) \geq 95\%]$ is equivalent to rejecting when it is in the outer right 5% tail of the distribution $[p(X \geq x) \leq 5\%]$. However, this is not necessarily true for discrete distributions (illustrated in figure 1). If the cumulative probability of an observation in the discrete binomial distribution lies above the 95% threshold, Wimmerstedt rejects this draw from the MC simulation. However, it is possible that the outer right tail probability of this observation is still more than 5% $[p(X \geq x) \geq 5\%]$. Due to this, in the case of the VaR-99.5%, Wimmerstedt implicitly uses a confidence level of $\alpha \approx 13\%$, which deviates heavily from the usual $\alpha = 5\%$ confidence level for rejection of a null hypothesis. This is an important factor in the relatively low supposed acceptance rate for Emmer's method of only 78%.

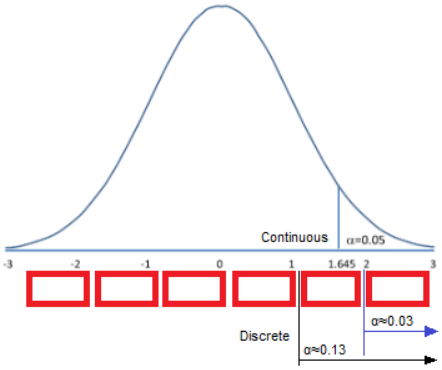

Figure 1: Continuous and discrete rejection regions; the bell curve represents a continuous function, whereas the red blocks represent a discrete distribution

We already mentioned that we exclusively use a one-tailed rejection region; Graham and Pál (2014)

explain the reasoning behind this. They state that a null hypothesis is rejected if the realised value of the employed test statistic $\hat{z}$ is significantly less than or significantly more than its assumed value under the null hypothesis, denoted z_0. However, only one of these instances (either $z >> z_0$ or $z << z_0$) represents underestimation of tail risk, while the other provides evidence for excessive coverage of the tail risk, through its overestimation. This last case indicates a conservative stance, and is therefore acceptable from a regulatory viewpoint. On the other hand, overestimation will lead to inefficient use of capital (Garcia-Jorcano, 2017), which is a concern for financial institutions.

Wong (2008) proposed using a saddle-point or small sample asymptotic technique for modelling of the tail distribution, which is relatively robust for small samples. Still, it is reliant on the assumption of normality. Graham and Pál (2014) expand upon this saddle-point technique, generalising it and using non-parametric methods for the estimation of distributions and theoretical acceptance/rejection thresholds, thus relaxing the normality assumption.

The method by M. Righi and Ceretta (2013) uses a dispersion of a truncated distribution by the estimated VaR upper limit, as opposed to a dispersion of the full distribution like in Wong's method. This method also allows for daily evaluation of ES predictions, without needing to wait for an entire backtesting period; and it is not limited to the normal distribution either, allowing the risk manager to choose the most appropriate distribution in a given scenario (M. B. Righi & Ceretta, 2015). This is a parametric bootstrapping method.

Acerbi and Szekely (2014) propose three backtest methodologies, which are rather comparable. Simply put, these methods test the likelihood that the observed shortfall originated from the distribution that was used for its prediction. This means that the actual ES prediction is of little consequence and its mostly about the assumptions that the prediction is based on. Since the test statistics that are given by Acerbi & Szekely dont follow any known distribution, they use Monte Carlo simulation for the evaluation of the significance of the test. This generation of a grid of quantiles from the cumulative density function (CDF) to compare to the observed exceedances is referred to as a parametric bootstrap. The first method uses a setup that assumes VaR to be correct, the second tests both ES and VaR jointly. The third method tests the tail observations by ranks, assuming these are uniformly distributed, which is reminiscent of the method by Emmer et al. (2015), where this is done in a discrete manner. Seeing how Wimmerstedt (2015) analysed all these methods and concluded that the first clearly had more power than the other two, we disregard the second and third methods in favour of the first, in order to keep the scope of this research manageable.

Del Brio et al. (2017) focus their research mainly on the optimisation of ES predictions, rather than their backtesting. As with Dalne's paper, however, this can still be of use to us, since it gives insight in what backtesting methods are generally accepted and used in the field. The backtesting method that is used here is a t-test based on VaR violation residuals. It seems comparable to a significance test where we assume standard normal residuals, but do not know the variance and thus must compensate for the fact that we use a sample standard error instead. Overall, this does not seem realistic, as already discussed

in the introductory paragraph, and the whole method is a step back in terms of both complexity and accuracy, compared to the more elegant methods discussed previously. We note that the test statistic is a very intuitive and simple risk measure to work with, given that it is the ES estimation error scaled by the estimated ES magnitude (distance from the total sample mean μ).

The main relevance of the paper by Garcia-Jorcano (2017) to us is that it contains an overview of what backtesting methods have been used lately. In particular, they comment on the method by Costanzino and Curran (2015), stating that 'it is appropriate for backtesting any spectral risk measure, including ES'; they also discuss the conditional test of Du and Escanciano (2016). Both of these methods are in essence continuous limit adaptions of the idea of Emmer et al., since they jointly evaluate a continuum of VaR levels. Furthermore, these two tests are two-sided and thus do not solely focus on risk underestimation but also reject the null-hypothesis in case of risk overestimation (Garcia-Jorcano, 2017).

Unfortunately, most of the concluding points of the paper focus on what forecasting methods perform best for the different backtesting methods, and thus the paper does not provide an objective evaluation or nominal measure of accuracy of an ES prediction. Thus, the topic of backtests of ES is not yet concluded. An interesting point, concerning Garcia-Jorcano's conclusion, is that the method of M. Righi and Ceretta (2013) and those of Acerbi and Szekely (2014) are more suitable for application on non-parametric ES forecasts, while the Graham and Pál (2014), Costanzino and Curran (2015) and Du and Escanciano (2016) methods are best used if we estimate ES by a parametric approach. Lastly, particular emphasis is laid on the difference between 1-day and 10-day ES estimation horizons, resulting in differing power levels for the prediction and backtesting sequences.

Bayer and Dimitriadis (2018) propose two regression-based backtests, which solely require ES forecasts as input parameters. No other quantities (e.g. VaR, volatility, tail distributions) are necessary for this method, as opposed to most previously mentioned backtesting procedures. Only one of these tests is one-sided, but equivalent in almost all other aspects, thus, we will focus on this one over the two-sided version. Its performance is compared to that of the Conditional Calibration backtest by Nolde, Ziegel et al. (2017), and they find that their own method is more powerful. To keep the scope of the research somewhat manageable, we will therefore not go into further detail on the methods by Nolde et al. (2017).

Löser et al. (2018) propose a so-called 'Unconditional Coverage' backtest of ES, which is unique in that it is appropriate in a multivariate setting. It is an extension to the method proposed by Du and Escanciano (2016), and compared to this, its main advantage is that the distribution is known for finite out-of-sample size (Löser et al., 2018). The researchers find that the proposed method clearly outperforms the alternative of Du and Escanciano (2016), though the test suffers from slight size distortions when using an in-sample period $T < 2500$.

Another method that is focused on Unconditional Coverage backtesting, by Moldenhauer and Pitera (2018), employs a so-called 'secured position' as risk measure, which gives us an indication of

whether the capital reserve of the financial product is sufficient to cover its risk. This method is compared to the second test by Acerbi and Szekely (2014), and is reported to give slightly better or similar results.

Furthermore, they suggest a traffic light system for ES, comparable to that which is currently in use in the VaR framework as stipulated by Basel regulation.

Finally, Moldenhauer and Pitera (2018) 'note that the regulator proposed the reference risk level change for VAR/ES migration in the similar fashion: ES-2.5%(X) is equal to 2.34, while VAR-1%(X) is equal to 2.33. This allows a smooth transition between VAR and ES frameworks (at least if the secured position distribution is close to normal).'

To conclude this part of the literature review, we have seen that there is quite a variety of flavours to choose from, when we wish to backtest ES predictions. Furthermore, many of the proposed methods build upon previous ones. Our task is to determine which of these are most appropriate to use in practice, from a regulatory perspective, foremost. We select the best-developed method per type of backtesting approach, and compare these to one another.

The method by Graham and Pál (2014) will be evaluated, because it expands upon the saddle-point method by Wong (2008). Löser et al. (2018)'s method is more refined than those of Costanzino and Curran (2015) and Du and Escanciano (2016), and the method by Moldenhauer and Pitera (2018) is based on the same type of approach, thus comparing these two methods by Löser et al. and Moldenhauer & Pitera respectively could give very valuable insights. The regression-based method (Bayer & Dimitriadis, 2018) approaches the problem from an entirely different angle, and will thus be considered as well. Del Brio et al. (2017) use a rather simple method, which might be used as a baseline reference point of how well the other methods perform. Finally, the method by M. Righi and Ceretta (2013) will be analysed, as per their own suggestion: 'controlled experiment approaches should be performed in order to compare size and power of SD based ES backtest with other approaches', which leads us to the section on how this analysis can be done, according to previous literature.

2.2 The analysis of a backtesting method

Dalne (2017) uses the method by Righi and Ceretta for the backtesting of ES estimates of market risk models, in order to select the most appropriate modelling distribution. This shows the practical application of Righi and Ceretta's method, though through this, the validity of Dalne's conclusion depends heavily upon the validity of Righi and Ceretta's method. Furthermore, this illustrates the inherent issue that all backtest results have: the uncertainty of whether the prediction method or the backtesting method was inappropriate (in case of rejection) or that either both were correct or both were incorrect (in case of acceptance).

The procedure of simulating data for the evaluation of relative power of backtesting methods is used by Wimmerstedt (2015). The paper documents rejection and acceptance rates of the previously mentioned methods of Wong (2008), M. Righi and Ceretta (2013), Acerbi and Szekely (2014) and Emmer et al. (2015), when applied to predictive models based on the standard normal distribution as well as

t-distributions with a variety of dof (representing more or less resemblance to the standard normal distribution). The true underlying distribution that is used for the simulation is known (standard normal), thus the researcher knows in which cases rejection is desirable and in which cases acceptance is.

In addition to providing us with a useful backtest, Löser et al. (2018) also examine its power by simulation, for the construction of a controllable but realistic scenario. They extend the setup of Du and Escanciano (2016), which is in turn a more elaborate version of what Wimmerstedt (2015) does. Two scenarios are considered; in the first, there is a structural break, and in the second the risk model is misspecified. Data generation is done in the first scenario using an AR(1)-CCC-GARCH(1,1) model with normal innovations before the break and t-distributed innovations afterwards. In the second scenario, a multivariate GARCH in mean model with normal innovations is used.

Bayer and Dimitriadis (2018) also use an MC method for determining both the size and the power of their backtest, along with that by McNeil and Frey (2000) and Nolde et al. (2017). They simulate an EGARCH(1,1) model with t-distributed innovations, with parameters calibrated using S&P 500 daily returns. Furthermore, they use a GARCH(1,1) model in a setting of continuous misspecification.

Lastly, an important part of backtest evaluations is size-adjusted power (Lloyd, 2005). This is a distortion of the true rejection rate of a backtesting method, normalised at $\alpha = 5\%$ when the underlying data truly follows the distribution of the null hypothesis. Through this, backtesting methods can be compared more fairly, despite having different sizes. For a detailed account of the procedure for constructing size-adjusted power, see Appendix Section A.1 or the paper by Lloyd (2005).

2.3 Statistical properties of risk measures and backtesting

Acerbi and Szekely (2017) state that backtesting in general remains to date a collection of disparate practices in the wait for a clear denition. They also comment on elicitability, stating that it serves to the purpose of conducting relative nondirectional tests of goodness between competing models issuing point predictions on a statistic. This type of procedure allows for model selection among multiple models rather than model validation, which needs an absolute scale for testing even a single model, and is the goal of a proper backtesting procedure. This suggests that the property of elicitability is irrelevant for proper backtesting, despite previous literature's insistence that this property is essential for the backtestability of a risk measure (Fissler, Ziegel & Gneiting, 2015).

Acerbi and Szekely (2017) also define the property of 'sharpness'. If a statistic is sharp, better predictions of the true underlying value give a better 'score' for this statistic than worse predictions. This is relevant for traffic light evaluations, since this gives a range of ordered values from better to worse, as opposed to a simple binary evaluation of a 'correct' or 'incorrect' prediction. An elicitable statistic is also sharp, but it does not necessarily provide directionality. It gives relative score, but not whether the model is over- or underestimated. Thus, a model that underestimates can be 'preferred' to one that overestimates 'slightly more heavily', which is not what we want for our research.

3 Methodology

In this section we will present the methods to be analysed, after which we will discuss the simulation scenarios that these methods will be tested in.

3.1 Backtesting methods

The following is a description with formulae of the backtesting methods that are analysed in this research. They are ordered according to their complexity, though admittedly, this is a rather subjective measure. Although all methods essentially test the null hypothesis of the ES being correctly specified, the methods perform this test through a variety of different mechanics. Therefore, we describe the immediate property that is tested as the method's specific null hypothesis.

3.1.1 Del Brio, Mora-Valencia and Perote

The least complex/elaborate test that we analyse employs the t-statistic (Del Brio et al., 2017). We test the returns against the predicted ES, scaled by the difference between the predicted ES and the sample mean. Only the returns that exceed the estimated VaR level are considered. Thus, this method assumes correct specification of the VaR level, and given this, tests the observed versus predicted loss, not the whole distribution.

First, the violation residuals X are calculated from the returns, or first differences of the P&L values, r_t:

$$x_t = \left(\frac{r_t - \hat{e}_t(\nu)}{\hat{e}_t(\nu) - \mu_t} \right) \mathbb{I}_{\{r_t < q_t(\nu)\}}, \tag{4}$$

where $\hat{e}_t(\nu)$ represents the estimated ES, $q_t(\nu)$ represents the VaR threshold under the null hypothesis and μ_t represents the conditional mean return over the past T observations, at time t. The test statistic is obtained in the following way (McNeil et al., 2005):

$$\hat{t} = \frac{\overline{X}}{S/\sqrt{\tau}}, \quad \tau = \sum_{t=1}^{T} \mathbb{I}_{\{r_t < q_t(\nu)\}}, \tag{5}$$

with $\overline{X}$ and S denoting the sample mean and standard deviation of the violation residuals x_t of $T = 252$ observations. The null hypothesis of zero mean violation residuals ($\overline{X} = 0$) is thus tested through a simple t-test with $\theta = \tau - 1$.

If the VaR level used in this methodology is misspecified, then this backtesting method could easily give a 'positive' evaluation of an incorrectly specified ES level too. This can be argued easily analytically as follows. The test statistic of Equation (5) is for the location of the mean of a subset of the observed data, specifically the VaR exceedances. Consider a hypothetical situation in which both the ES and VaR are misspecified, in the direction of underestimation of the tail risk, as illustrated in Figure 2. This misspecification creates a situation where more than just the lowest 2.5% of observations are counted as an exceedance to be used for the location of mean test, because the hypothesised VaR threshold is less strict than the true VaR threshold. Due to this usage of an incorrect subset of the data, the 'observed

mean' of the tail subset will always be greater than or equal to the true ES value of the data, causing underestimation of the tail risk. Following this, the 'observed ES', which is a distortion of the true value, will lie between the predicted ES and the true ES of the dataset. Through this mechanic, the VaR misspecification will cause the test to give a much smaller p-value than it should, causing the rejection rate to be far too low. We can conclude from this that it is necessary to perform a test on the VaR level before or in parallel to this testing procedure, in order to evaluate the reliability of the results from this test.

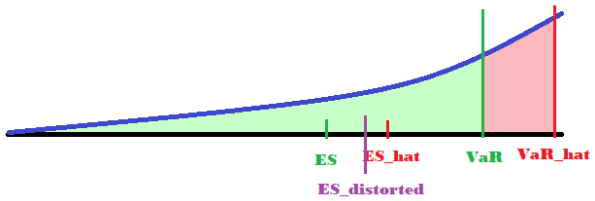

Figure 2: Depiction of left tail of an arbitrary distribution, with the true ES and VaR levels (green), misspecified values (red) and numerical mean of exceedances of the misspecified VaR level VaR_hat (purple)

3.1.2 Righi and Ceretta

The backtesting method by M. Righi and Ceretta (2013) uses a dispersion of a truncated distribution by the estimated VaR upper limit. It performs a one-tailed test with the null hypothesis being that the most extreme 2.5% tail losses behave as specified by the assumed distributional form; the alternative hypothesis being that the occurred tail losses of the specified level are more excessive than predicted according to the specified prior distribution. Through this methodology, all observations are compared to the predicted ES, scaled by the variance of the tail as per the specification of the assumed distribution. Thus, the entire tail distribution under the null hypothesis is analysed. Because of this, the method does not require a prior VaR backtest to be valid.

The backtest statistic is defined as follows:

$$BT_t = \frac{r_t - \hat{e}_t(\nu)}{SD_t(\nu)}, \tag{6}$$

where $SD_{t+k}(\nu)$ represents the Shortfall Deviation, the standard deviation of any observation exceeding the specified VaR level ν. This formula can be written out more explicitly as follows:

$$BT_t = \frac{r_t - E_t[r_t | r_t < F^{-1}(\nu|\pi_t)]}{(Var_t[r_t | r_t < F^{-1}(\nu|\pi_t)])^{1/2}}, \tag{7}$$

with $F^{-1}(\nu|\pi_t)$ representing the analytical VaR threshold and π_t the distribution parameters under the assumed distribution of the null hypothesis, and Var representing the variance; giving us a test statistic that 'normalises' VaR exceedances r_t through subtraction of ES and division by SD of the corresponding time period, under H_0. The normalised statistic is compared to a grid of random draws ($n = 10^5$)

from the distribution under the null hypothesis (of which only VaR threshold exceedances are taken into account and normalised), according to the original methodology by Righi & Ceretta, to analyse the likelihood of observing the given exceedance under the null hypothesis. We calculate the percentage of values of the grid of draws n that is lower than BT_t as p_t^*. This MC sampling is done $N = 10^3$ times, resulting in N values of p_t^*; we take the median of these values p_t^* as the p-value of the backtest. The MC algorithm for this method can be seen in M. Righi and Ceretta (2013).

This methodology backtests individual VaR exceedances, which is notably different from most backtests that test a whole year's worth of observations at once. The results must thus be adapted for the comparison between this method and the others. For this adaptation, we simulate $N = 10^5$ yearly scenarios of $T = 252$ random draws, instead of the $N = 10^3$ samples of $n = 10^5$ draws. We then take the aggregate of the simulated BT values for the entire year, which gives a range of N values as per an MC simulation. Next, we simply take the $\alpha * N$-th order statistic as our threshold value to compare the observed aggregated BT to, to give a yearly evaluation in favour of daily evaluations for all exceedances.

3.1.3 Moldenhauer and Pitera

As mentioned in the literature section, the method by Moldenhauer and Pitera (2018) employs a secured position. The secured position is constructed as the difference between the risk estimator ES ($\hat{e}_t$) and the P&L value at time t (r_t), proportional to portfolio volatility:

$$x_t = \frac{r_t - \hat{e}_t(\nu)}{-\hat{e}_t(\nu)} = 1 - \frac{r_t}{\hat{e}_t(\nu)}, \quad \text{where } \hat{e}_t(\nu) < 0 \ \forall \ t. \tag{8}$$

Under the null hypothesis the sum of the lowest 2.5% secured positions is equal to 0:

$$H_0 : \quad \sum_{i=1}^{\lfloor \nu T \rfloor} x_{(i)} = 0, \tag{9}$$

with $x_{(i)}$ representing the i-th order statistic of $x_1, ..., x_T$. The test statistic that follows from this is constructed in the following way:

$$G_T = \sum_{i=1}^{T} \frac{\mathbb{I}_{\{x_{(1)} + ... + x_{(i)} < 0\}}}{T}. \tag{10}$$

What this represents in real terms is a count of how many of the most severe loss observations it takes for their sum to be greater than the sum of their corresponding ES predictions. Though it might seem a bit convoluted at first, it is actually a very intuitive measure of the ES. However, it calculates what percentile of the observed sample corresponds with the given ES prediction, instead of analysing the realised ES of the 97.5th percentile of the trading year. In short, we measure the sum of the biggest number of worst P&L realisations that is still more negative than the sum of their corresponding ES predictions. The implicit null hypothesis is that the percentile of the observed returns that gives the same ES value as what was predicted is smaller than or equal to ν, or in terms of the test statistic: $G_T \leq \nu$.

The test statistic can additionally be transformed to match the original VaR framework and its traffic light scheme. We use a transformation to nominal values: $X = T * G_T = 252 * 0.025 \approx 6$

observations under the null hypothesis. The new traffic light system, given in table 1, gives the probability levels and 'danger zones' corresponding to the nominal X values and G_T intervals under normality.

zone	observations	probability	G_T
green	$X < 12$	90%	[0.00, 0.05)
yellow	$12 \leq X < 25$	9.99%	[0.05, 0.10)
red	$25 \leq X$	0.01%	[0.10, 1.00]

Table 1: Traffic lights of Moldenhauer and Pitera's method under normality

In order to compare this method to the others, however, we will need to get the 95% confidence level. Thus, we need the exact value of X (or G_T) that this corresponds to as rejection region. This region depends on the underlying assumptions, though, and can be approximated using MC. For the standard normal distribution, we know that $ES_{2.5\%} = -2.34$ (Moldenhauer & Pitera, 2018). What we can do is take random samples from the normal distribution (50000 iterations) and analyse what value of X (or G_T) we get for the given ES value in the 5% most extreme cases, for the comparability with the other methods on a significance level of $\alpha = 5\%$. Whenever considering another distribution, this MC method must be run once more, using draws from the corresponding distribution and the corresponding ES value.

This method does not need any VaR prediction input and can thus be performed completely independently from any VaR evaluation.

3.1.4 Bayer and Dimitriadis

The regression-based method by Bayer and Dimitriadis (2018) uses a joint regression on both ES and VaR, though the VaR regression part is not strictly necessary for the one-sided test. The joint regression framework for semi-parametric estimation is stated as follows:

$$r_t = \beta_0^q + \beta_1^q \hat{e}_t(\nu) + u_t^q, \tag{11}$$

$$r_t = \beta_0^e + \beta_1^e \hat{e}_t(\nu) + u_t^e, \tag{12}$$

using the same notation as before. The following holds for the error terms: $VaR_\nu(u_t^q|\mathcal{I}_{t-1}) = 0$ and $ES_\nu(u_t^e|\mathcal{I}_{t-1}) = 0$. This also gives us:

$$ES_\nu(r_t|\mathcal{I}_{t-1}) = \beta_0^e + \beta_1^e \hat{e}_t(\nu). \tag{13}$$

For the one-sided Intercept ESR Backtest, we fix $\beta_1 = 1$, giving us the following regression equation:

$$r_t - \hat{e}_t = \beta_0^e + u_t^e. \tag{14}$$

We test whether parameter β_0^e equals zero. The one-sided hypothesis is formulated as follows:

$$\mathbb{H}_0^{1s} : \beta_0^e \geq 0 \quad \text{against} \quad \mathbb{H}_1^{1s} : \beta_0^e < 0, \tag{15}$$

on which we perform a t-test, based on an asymptotic covariance and bootstrap procedure. The t-statistic for this Intercept ESR Backtest is given as follows:

$$t_I = \frac{\hat{\beta}_0^e}{\sqrt{\hat{\Sigma}_{22}/T}}, \tag{16}$$

where Σ_{22} represents the bottom right element of the asymptotic covariance matrix, given in the joint estimation of the sample Quantile and ES (Dimitriadis & Bayer, 2017), constructed by regressing the observed losses on a constant only. The formulae for the elements of Σ are given as follows:

$$\Sigma_{11} = \frac{\nu(1-\nu)}{f_R^2(Q(\nu))}, \tag{17}$$

$$\Sigma_{12} = \Sigma_{21} = (1-\nu)\frac{q(\nu) - \hat{E}(\nu)}{f_R(Q(\nu))}, \tag{18}$$

$$\Sigma_{22} = \frac{1}{\nu}\mathrm{Var}(R - Q(\nu)|R \leq Q(\nu)) + \frac{(1-\nu)}{\nu}(Q(\nu) - \hat{E}(\nu))^2, \tag{19}$$

where $Q(\nu)$ and $\hat{E}(\nu)$ represent the VaR and ES under the null hypothesis in vector notation and R represents the vector $[r_1 \ ... \ r_T]$. As can be seen in Equation (16), we only need the bottom right element of the covariance matrix, since this is the ES part, whereas Equation (17) concerns the Quantile part of the formulation and the off-diagonal elements of Equation (18) are concerned with covariance, of course. Therefore, we will not go into detail on f_R.

A paired bootstrap procedure is used for the construction of the CI (Efron & Tibshirani, 1994), since neither the loss function of the M-estimator, nor the asymptotic covariance depend on the temporal ordering of pairs $(r_t, \hat{e}_t)$. We take $M = 10^3$ bootstrap samples from the errors $u_t^e = r_t - \hat{e}_t(\nu)$ as $U^{(b)}$. The t-statistic is calculated as in (22), centered around the $\hat{\beta}_0^e$ value:

$$t^{(b)} = \frac{ES_\nu(U^{(b)}) - \hat{\beta}_0^e}{\sqrt{\hat{\Sigma}_{22}^{(b)}/T}}, \tag{20}$$

where $\hat{\Sigma}^{(b)}$, of course, represents the asymptotic covariance matrix estimate of bootstrap set $U^{(b)}$. This gives us an evaluation of how unlikely the original t_I was; we reject the null hypothesis if $t_I < t_{(\alpha M)}^{(b)}$, with $t_{(\alpha M)}^{(b)}$ representing the $\alpha * M$-th order statistic of sample $t^{(b)}$.

This method uses the VaR level under the null hypothesis as input for calculation of Σ_{22}, and is thus dependent on the assumption of a correctly specified VaR level, as the DMP method in Section 3.1.1. his VaR-dependency issue can be lessened by using the empirical VaR value of the most recent set of observations instead of basing the VaR off the same prediction method as the ES is. However, performing a VaR backtest is advisable in any case. This goes for all the VaR-dependent methods.

3.1.5 Graham and Pál

The backtesting method by Graham and Pál (2014) expands upon the Lugannani-Rice approach by Wong (2008) and generalises it. A small-sample asymptotic saddle-point technique is used, and made

analytically tractable and operationally feasible. Simply put, the original method by Wong is made more implementation-friendly.

Per exceedance of the VaR level, as specified under the null hypothesis, we calculate statistic x_t. There are several procedures for defining x_t, namely via historical simulation (using an empirical CDF), variance-covariance (fitting a standard known distribution) or MC simulation. This choice of CDF is where we see the null hypothesis being tested, since this methodology essentially evaluates the fit of this specification for the tail distribution on the observed data.

The simplest way of specifying x_t is using the empirical distribution:

$$x_t^e = \ln\left(\frac{\sum_{i=1}^{N} \mathbb{I}_{\{r_{t-i} \leq r_t\}}}{\nu N}\right), \tag{21}$$

where N represents the historical sample size used, which is set equal to T.

When using a standard distributional CDF, such as the normal or t-distribution, we use the following:

$$x_t^s = (\ln\, F_t(r_t|\pi_t) - \ln\, \nu)\mathbb{I}_{\{\ln\, F_t(r_t|\pi_t) < \ln\, \nu\}}, \tag{22}$$

with $F_t(\cdot)$ being the percentile value of an observation within its forecast distribution (CDF), that is, the prior assumed distribution that the predictions are based on.

Finally, an EVT tail model, in this case the GPD, can be used, giving us:

$$x_t^{EVT} = \begin{cases} -\frac{1}{\xi}\ln\left(1 + \frac{\xi}{1-\xi}\frac{q_t(\nu) - r_t}{r_t(\nu) - \hat{e}_t(\nu)}\right), & \text{if} \quad 0 < \xi < 1, \\ \frac{r_t - q_t(\nu)}{q_t(\nu) - \hat{e}_t(\nu)}, & \text{if} \quad \xi = 0. \end{cases} \tag{23}$$

We tend to use distributional assumptions in our analyses, thus we generally apply Equation (22) in this research, giving $x_t = x_t^s$. If we use this distributional CDF approach, we evaluate the entire hypothesised tail distribution as was the case with Righi & Ceretta's method of Section 3.1.2. Thus, again, we do not strictly require a VaR backtest.

We take the mean of the statistics x_t as $\overline{X}$, in order to solve the saddle-point equation:

$$K'(s) = \frac{M'(s)}{M(s)} = -\frac{\nu}{(s+1)[s(1-\nu)+1]} = \overline{X} \quad \text{for} \quad \overline{X} < 0. \tag{24}$$

This gives the unique solution $\tilde{s}$ on interval $(-1, \infty)$:

$$\tilde{s} = \frac{(\nu - 2) + \sqrt{\Delta}}{2(1-\nu)} \quad \text{for } \Delta = \nu^2 + \frac{4\nu(\nu-1)}{\overline{X}}. \tag{25}$$

We define the following variables for the construction of the p-value:

$$\eta = -s\overline{X}\sqrt[4]{\Delta}\frac{\sqrt{\nu T}}{\nu} \quad \text{and} \quad \varsigma = \text{sgn}(s)\sqrt{2T(s\overline{X} - K(s))}, \tag{26}$$

where $\text{sgn}(s)$ takes the sign of s if $s \neq 0$, and is 0 otherwise, and $K(s) = \ln(\frac{\nu}{s+1} + 1 - \nu)$. We can simplify the formulae of Graham and Pál (2014) page 69 for $K(\cdot)$'s second and third order derivatives as:

$$K''(0) = \nu(2-\nu), \quad \text{and} \quad K'''(0) = 2\nu[(1-\nu) - (2-\nu)^2], \tag{27}$$

since we have no need for the general forms $K''(t)$ and $K'''(t)$.

The calculation of the Lugannani-Rice formula is done next, resulting in the p-value under the null hypothesis:

$$\hat{p} = \mathbb{P}[\overline{x} \leq \overline{X}] = \Phi(\varsigma) - \phi(\varsigma)\left(\frac{1}{\eta} - \frac{1}{\varsigma}\right) \quad \text{for} \quad \overline{X} \neq -\nu, \tag{28}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ represent the standard normal CDF and PDF, respectively.

In case that $\overline{X} = -\nu$, the p-value is calculated as follows:

$$\hat{p} = \mathbb{P}[\overline{x} \leq \overline{X}] = \frac{1}{2} + \frac{K'''(0)}{6\sqrt{2\pi T[K''(0)]^3}}. \tag{29}$$

This is highly uncommon in practice but will be necessary to define, given our tendency for large simulations.

3.1.6 Löser, Wied and Ziggel

The LWZ backtest is based on the cumulative violation process; the sum of all VaR exceedances in a given period. Through this, the method backtests VaR and ES jointly as a ridge backtest (Acerbi & Szekely, 2017). Due to a transformation of the tail observations we get *i.i.d.* uniform random variables (under the null hypothesis), the sum of which gives us the Irwin-Hall distribution (Hall, 1927). This gives us the test statistic S_{UC} in Equation (30), which is uniformly distributed from 0 to 1 under the null hypothesis when T tends to infinity, thus giving the rejection region [0, 0.05).

$$S_{UC} := \frac{1}{1-\nu^T} \sum_{i=1}^{T} \binom{T}{i} \nu^i (1-\nu)^{T-i} \Upsilon_i(\hat{H}_{\cdot n}), \tag{30}$$

where $\hat{H}_{\cdot n}$ represents the cumulative violation process for n observable trading days:

$$\hat{H}_{\cdot T} = \frac{1}{\nu} \sum_{t=1}^{T} (\nu - F_{t|t-1}(r_t|\pi_t)) \mathbb{I}_{\{r_t < F^{-1}_{t|t-1}(\nu|\pi_t\}}. \tag{31}$$

$\Upsilon_i(\cdot)$ represents the CDF of the Irwin-Hall distribution:

$$\Upsilon_i(x) := \frac{1}{i!} \sum_{j=0}^{\lfloor x \rfloor} (-1)^j \binom{i}{j} (x-j)^i, \tag{32}$$

for which we use a normal approximation due to mathematical instability:

$$\Upsilon_i(x) \approx \Phi\left(\frac{x - i/2}{\sqrt{i/12}}\right). \tag{33}$$

The cumulative violation process $\hat{H}_{\cdot T}$ is dependent on the prior distributional assumption that the ES predictions are based on, the validity of which is essentially what this procedure evaluates.

One could note the questionability of the assumption of T tending towards infinity. However, $T = 252$, the amount of trading days that we employ throughout the research, is sufficient.

3.2 Simulation analyses

Several analyses of the proposed methods are to be performed. These comparisons of the methods make use of simulated data, which can easily be controlled to behave nicely, such that we can have a clear evaluation of the performance of the backtest in question. There are four different scenarios that we will be researching, the normal- and t-distribution setup by Wimmerstedt (2015), the CCC-GARCH setup by Löser et al. (2018), the GARCH setup by Bayer and Dimitriadis (2018) and a simple portfolio setup employing the Variance-Covariance method for VaR and ES prediction.

3.2.1 Wimmerstedt

The methodology of Wimmerstedt (2015)'s fourth and fifth chapters allows for a comparison of the acceptance rates of the suggested methods in a setting where the ES predictions are specified to be correct (meaning they are equal to the analytical expectation under the specification of the simulated data), and the rejection rates in a setting where we know the predictions are incorrect. We use a combination of the assumed and true tail distributions being equal to the standard normal or t_3-distribution, giving us the 4 scenarios of correctly assumed normal (i), incorrectly assumed normal with true t_3 (ii), incorrectly assumed t_3 with true normal (iii) and correctly assumed t_3 (iv). We expect a theoretically ideal backtest to give 95% acceptance for both correct specifications (i) and (iv) and a very low acceptance rate for the incorrect specification (ii). We expect an acceptance rate above 95% for incorrect specification (iii), since we perform a one-sided backtest on underestimation of tail risk.

Whenever a t_3-distributed tail is used, as is described in Wimmerstedt (2015), only the losses in excess of the VaR quantile are drawn from a t_3-distribution, while the rest of the innovations are drawn from the (standard) normal distribution. The reason for this construction is that we do not get a disproportionate amount of VaR exceedances this way, but only an increase in the size of any one exceedance. Thus, we control for the backtesting methods' sensitivity to high numbers of exceedances and only test the capacity for detection of misspecification of the magnitude of the ES.

3.2.2 Löser, Wied and Ziggel

The setup by Löser et al. (2018), as discussed in their section on simulation, will be applied too. This entails a (univariate) CCC-GARCH(1,1) model with a structural break. The innovations ϵ_t will be drawn from a t-distribution with $\theta = \infty$ up until the structural break point N, giving Gaussian white noise. From period $N + 1$ onward, however, the innovations will be drawn from a t-distribution with $\theta \in \{\infty, 30, 22, 15, 10, 7, 5, 3\}$. We expect this to give a range of rejection probabilities per method, such that we can see how sensitive each method is to an increase in a 'fatness' of the tail losses. The model is specified as follows:

$$r_t = \rho r_{t-1} + v_t,\ v_t = \sigma_t \epsilon_t,\ \epsilon_t \sim t(\theta), \tag{34}$$

$$\sigma_t^2 = \omega + \alpha v_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{35}$$

with $|\rho| < 1, \omega \geq c > 0, \alpha, \beta \geq 0$ for some constant c. This method is also more of a stepping stone for the more elaborate simulation of Section 3.2.3.

3.2.3 Bayer and Dimitriadis

We also employ the simulation specifications by Bayer and Dimitriadis (2018). Again, a GARCH(1,1) model is used, this time following a slightly different specification:

$$r_t = \sigma_t z_t, \ z_t \sim t(\theta), \tag{36}$$

$$\sigma_t^2 = \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \tag{37}$$

with $\gamma_0 = 0.01, \gamma_1 = 0.1, \gamma_2 = 0.85$, and $\theta = 5$ for the true model. Five misspecification designs are defined, where the dgp is kept constant, but the assumed model is changed:

1. Changing how conditional variance reacts to squared returns, $\tilde{\gamma}_1 \in \{0.02, 0.2\}$, with $\tilde{\gamma}_2 = 0.95 - \tilde{\gamma}_1$ such that the process remains constant.
2. Changing unconditional variance to 0.4 - 0.01 by changing $\tilde{\gamma}_0$.
3. Changing persistence in shocks to 0.90 - 0.999 by setting $\tilde{\gamma}_1 = c\gamma_1$ and $\tilde{\gamma}_2 = c\gamma_2$ for a constant c, and $\tilde{\gamma}_0 = \mathbb{E}[\sigma_t^2](1 - \tilde{\gamma}_1 - \tilde{\gamma}_2)$ to keep the unconditional variance constant.
4. Increasing θ.
5. Increasing the threshold level ν to 5%.

Although item (5) will give a less extreme prediction of ES than the base scenario, which is the type of misspecification that we are looking for, it is not a realistic problem setting that we are concerned with. Bayer and Dimitriadis (2018) argue that it is a scenario of human error, when a forecaster submits predictions for some incorrect level of $\tilde{\nu} \neq \nu$. We research misspecifications in tail risk predictions and thereby methods for detection of mismatches in assumed and observed error distributions, not manual input errors for risk levels. Thus, we deem scenario (5) insignificant and beyond the scope of our research. Scenario (4) will also not be considered, due to its similarity to the analysis of Section 3.2.2.

For this analysis, we will compare the methods' size-adjusted power, since the sizes of the tests are rather far apart. This will give us a much fairer evaluation than a naive power comparison (Lloyd, 2005). We refer to Appendix Section A.1 for some comments on size-adjusted power.

3.2.4 Variance-Covariance anaylsis

The final analysis we perform is a step back in terms of statistical complexity. Since one of the most common methods for constructing a VaR prediction currently in use is the Variance-Covariance method, which approximates the VaR level of a portfolio through the statistics of the individual stocks, we construct a 2-stock portfolio from a bivariate normal distribution. Although the combination of two normal distributions results in a normal distribution, we wish to have two separate distributions of which we can manipulate parameter inputs, in order to research the effects of these misspecifications on the Variance-Covariance approximation of the lower quantile risk estimate. Precisely because the

method is an approximation and not necessarily the true analytical value, this allows us to compare our ES backtesting methods more realistically to the currently used methodology for VaR estimation and backtesting.

We use the following parameters under the null hypothesis, which the predicted ES and VaR values are based on:

$$R = w_1 X_1 + w_2 X_2, \quad \text{with } w_1 = 0.6,\ w_2 = 1 - w_1; \tag{38}$$

$$X \sim N(\mu, \Sigma), \quad \mu = [0, 0], \quad \sigma_1 = 0.07,\ \sigma_2 = 0.04,\ \rho = 0.3 \text{ and } \sigma_{12} = \rho\sigma_1\sigma_2, \tag{39}$$

The VaR and ES predictions are constructed by use of the Variance-Covariance method. This method is not perfect, but it is widely used; besides, the prediction of VaR and ES is not the main focus of this research, and thus lies outside our scope, so we will regard the Variance-Covariance method as adequate in this setup. We construct the predictions as follows:

$$\text{VaR}_{0.01} = \textit{investment} * z_{0.01} \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12}}, \tag{40}$$

$$\text{ES}_{0.025} = \textit{investment} * z_{0.025}^{ES} \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12}}, \tag{41}$$

$$\text{VaR}_{0.025} = \textit{investment} * z_{0.025} \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12}}, \tag{42}$$

where *investment* is normalised to unity, for simplicity. The z-scores are taken from the standard normal table: $z_{0.01} = -z_{0.99} = -2.33$ and $z_{0.025} = z_{0.975} = -1.96$. For the z-score for Equation (41) we use the analytical 2.5% expected shortfall value of the standard normal distribution: $z_{0.025}^{ES} = ES_{2.5\%} = -2.34$ (Moldenhauer & Pitera, 2018). $\text{VaR}_{0.01}$ will be used for the VaR backtest, since that is the threshold level that the VaR predictions have always had to be reported for, whereas $\text{ES}_{0.025}$ is obviously what the backtests will have to be performed on, as stated in the FRTB (BCBS, 2016). Furthermore, some of the backtests need the $\text{VaR}_{0.025}$ value as input, which is why we calculate this threshold value too.

The VaR backtest that is applied in this setup is as simple as they come, we reject the prediction if the observed amount of exceedances exceeds the predetermined threshold. For the appropriate size-adjusted power in this discrete setup with $T = 252$ trading days, we need a mixture of the thresholds of rejection at 5 or more exceedances (which gives a size of about 9%) and rejection at 6 or more exceedances (which gives a size of about 4%). More on this in Appendix Section A.1.

For the scenarios of misspecification, we have 3 categories:

1. Changing the correlation coefficient $\rho \in \{0, 1\}$.
2. Changing the standard deviation of one of the stocks $\sigma_2 \in \{0.02, 0.08\}$.
3. Changing the mean of one of the stocks $\mu_1 \in \{-0.14, 0.105\}$.

We start by changing ρ, since this is expected to increase the overall variance of the system. Thus, we analyse how variance increases affect rejection rates without touching the shape of the tail as is done in previous analyses, distorting the normal distribution into a t-distribution. The change in σ_2 accomplishes

much the same goal, and is there mostly to check for consistency. Finally, the change in mean makes for an analysis on the correctness of the location of the distribution, and thereby of the tail. What is interesting about this is that the ES risk measure might be influenced by this differently from the VaR risk measure. Compared to an increase in variance that distorts the bell curve (stretches it horizontally), the change in location simply moves the bell curve horizontally. This will also make for underestimation of the tail risk, but in a different way from how an increase in variance does. Therefore, we expect our methods to react in a different way from all the other misspecifications discussed in previous sections.

Contrary to the way the Bayer Analysis 3.2.3 is setup, we do not change the assumed distribution this time, but instead the underlying data generating process. Because of this, the size-adjusted power is completely accurate, whereas in the Bayer Analysis 3.2.3, this is not strictly the case, due to time constraints and the manner in which one adjusts power for size. More details on this are in Appendix Section A.1.

4 Results

The results of the analyses of the proposed backtesting methods will be discussed in this section. In Section 4.1 we discuss the results of the methods in a setup that is comparable to that of Wimmerstedt (2015)'s fourth and fifth chapter. In Section 4.2 we use the (univariate) GARCH structural break setup as in Löser et al. (2018)'s chapter 3 on simulation. In Section 4.3 we report on the GARCH setup as in Bayer and Dimitriadis (2018) and we discuss the results of the Variance-Covariance setup in Section 4.4. Finally, we report on the computational time and complexity in Section 4.5.

Throughout this section, either acceptance or rejection rates will be reported for all performed analyses, depending on what makes the most sense intuitively. Since the sum of acceptance and rejection rates of any analysis always equals 1, we need only report one of the two. In the 'size-adjusted power' analyses, rejection rate (which is equivalent to power) is obviously what we report. Generally, though, we report acceptance rates in correctly specified scenarios and rejection rates in incorrectly specified scenarios.

4.1 Wimmerstedt simulation

Here, we report on the preliminary analysis of the methods by the setup of Wimmerstedt (2015). In table 2, we show the acceptance rates of a correctly specified normally distributed setting (left column) and rejection rates of a t_3-distributed tail, when the assumed distribution is normal.

All methods have good acceptance rates of the true normal scenario, though BD seems to be too lenient, with an acceptance rate closer to 100% than the expected 95% (which is theoretically correct). Furthermore, LWZ is rather on the strict side with its acceptance rate of only 89%. This is particularly strange, given that the method is far too lenient in terms of rejection of the false scenario, with a rate of 47%. We would expect any method's bias to be in the same direction, either too lenient for all

specifications, or too strict for all, irrespective of what model assumptions are used. The rejection rate of DMP also seems to be rather low, with rejection only in 79% of all simulations under misspecification. The BD method is the worst-performing method in this setting, achieving a rejection rate of only 39%. We will go into more detail on possible explanations of this later.

Appendix tables 5, 6, 7, 8, 9 and 10 show all the results of the analysis, with acceptance rates under the four possible scenario combinations of normal or t-distributional assumptions and realisations. We observe that all methods have higher or equal acceptance rates for the correctly specified t-distributed tail with $\theta = 3$ than for the correctly specified normal tail, suggesting that these methods perform better under heavy-tailed specifications. Lastly, we observe an acceptance rate of 100% in all cases of overestimation of risk, as expected in the one-tailed testing scenario.

Method	Acceptance when true	Rejection when false
DMP	0.9619	0.7934
RC	0.9589	0.9775
GP	0.9531	0.9777
LWZ	0.8878	0.4742
MP	0.9570	0.9325
BD	0.990	0.390

Table 2: Acceptance/rejection rates per backtesting model (for a random number of exceedances); the true scenario being $N(0, 1)$, with t_3 being used for the false scenario. In order, the methods are those of sections 3.1.1 (DMP), 3.1.2 (RC), 3.1.5 (GP), 3.1.6 (LWZ), 3.1.3 (MP) and 3.1.4 (BD).

For the Bayer & Dimitriadis method we use $B = 10^3$ bootstrap iterations, as per their suggestion. We use $N = 10^4$ MC iterations, such that it takes about seven minutes (in the true normal case) for the python script to run. Larger simulations are not completely unfeasible, if one is willing to wait for an hour or more, and an important note here is that this does not reflect the time that the method normally takes to run, since the N simulations are only necessary for an overall acceptance rate, but not for the model evaluation of a single real data set. Furthermore, increasing these iteration amounts does not influence the acceptance/rejection rates much, only changing the fourth and sometimes third decimals of table 2.

When looking at the 'Rejection when false' column, a 'good' backtest should have a value as close to 1 as possible, representing 100% rejection under misspecification (power property). The results for the BD method, however, are far under par. Given that the rejection rate in the misspecified setting is lower than 40%, we can conclude that this method is far too lenient in the given setup. This is also supported by the acceptance rate, which would ideally be 95% (size property), but is much higher than that. One explanation for this is that bootstraps of a heavy-tailed sample give large biases in their test statistics. This hypothesis can be researched by using new simulations in favour of the bootstrapped

sets[5]. If these consistently give better results, that would indicate the existence of a bias in resampling when a small amount of extreme left tail losses is concerned. Furthermore, if we compute the t-statistics of the bootstrap samples using the variance of the original sample instead of the bootstrapped sample, we get better (and more reliable) results:

$$t_{alt}^{(b)} = \frac{ES_\nu(U^{(b)}) - \hat{\beta}_0^e}{\sqrt{\hat{\Sigma}_{22}/T}}. \tag{43}$$

This approach and the original are equal if $\hat{\Sigma}_{22}^{(b)} = \hat{\Sigma}_{22}$. This is very unlikely to occur, since bootstrap sample variances are much less stable than the original, due to the way the draws are done with replacements. As it turns out, this change in variance increases the rejection threshold, which results in a stricter acceptance criterion, which leads me to recommend it over the original. The rejection rate when false goes up to 0.7222, which is still not on par with the other methods, but it is an improvement on the original.

We conclude that in this specific setup the RC, GP and MP methods outperform the others, with the DMP method trailing closely. The LWZ and BD methods are both inaccurate in the true scenario, and have very low power in the false scenario. Thus, we conclude that these last two methods are inadequate at identifying incorrect specifications in this setup.

4.2 Löser, Wied and Ziggel simulation

The main methods of interest are analysed here in a similar manner to the previous section, this time using the simulation setup by Löser et al. (2018). This analysis can be considered visually by plotting θ against the resulting acceptance rate. The results are visible in Figure 3. One thing to note is the x-axis, which is of irregular step length. Due to this the curvature gives a biased view, however, we are primarily interested in the relative performance of the methods, which is very clear from this figure. We see an irregularity for the GP method when $\theta = 3$, in green in Figure 3, which should obviously converge to 0% acceptance for low dof instead.

We see that the RC and DMP methods are the least sensitive and the BD method starts out rather insensitive, but has a steeper slope around $\theta = 10$ compared to the other methods. The remaining three methods of LWZ, MP and GP have very similar performances and are the preferred methods when we consider this analysis in isolation.

We see that the GP method is the quickest to react to the increased fat-tailedness through the lowering of θ, in addition to it being the strictest method in its acceptance overall. One note for this method is that it seems to break when the error terms behave too erradically. This is the case in the setup with $\theta = 3$, where the method gave an acceptance rate of over 40%, even though the less inaccurate specification of $\theta = 5$ resulted in a much lower acceptance rate. This is a very extreme situation, though, and can easily be noticed and adjusted. On the other hand, if this method performs identically compared to others overall, then this flaw would make for the choice against the GP method.

[5]This is outside the scope of our research, and thus we suggest it for future research.

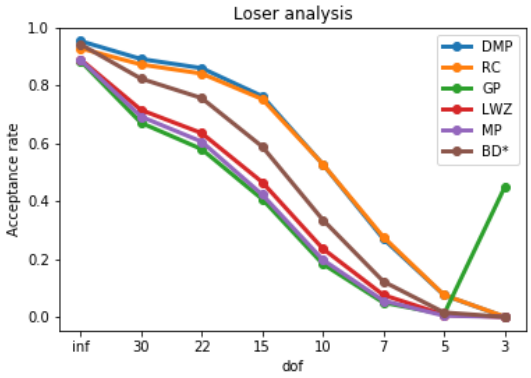

Figure 3: Acceptance rate per backtesting model plotted against θ, based on the GARCH structural break setup (univariate)

When considering the BD methodology, we observe better size and power properties for the adjusted version (BD*) compared to the original method for all levels of θ (see appendix Table 11). This result, in addition to what was discussed in Section 4.1, leads us to employ the adjusted method for all subsequent analyses instead of the original, abandoning the original altogether. Besides the observed favourability of the adjusted method, we also have theoretical reasoning for this choice in the fact that the bootstrapped variance is very unstable, giving much less reliable results compared to the original sample variance.

4.3 Bayer and Dimitriadis simulation

Here we present the results of the simulation from the setup by Bayer and Dimitriadis (2018). We simulated the first three sets of scenario changes: increase in reaction to squared returns (Figure 8), decrease of unconditional variance (Figure 9) and increase in persistence in shocks (Figure 10).

4.3.1 Impact on conditional variance estimates

In order to understand the impact on acceptance rates of the backtesting methods, we must first understand the impact of the misspecification on the conditional variance predictions. The true values of the conditional standard deviation σ_t are given in Figure 4.

Figures 5, 6 and 7 show true values of the conditional standard deviation σ_t (orange) versus predictions $\hat{\sigma}_t$ under the correct specification (blue dots), where the backtest should be rejected for underestimation of risk according to Bayer and Dimitriadis (2018) (red line) and where the backtest should be accepted for overestimation of risk according to Bayer and Dimitriadis (2018) (green line).

All graphs have been shown for the same period in the simulation, zoomed in for visibility; the behaviour is consistent through the entire simulation period. What we expect based on the research of Bayer and Dimitriadis is that the red lines are on average below the orange line, whereas the green lines should be above it. The only situation where this is true is for the Unconditional variance change, Figure 6, which is why that analysis is the only one which is consistent with that of Bayer and Dimitriadis

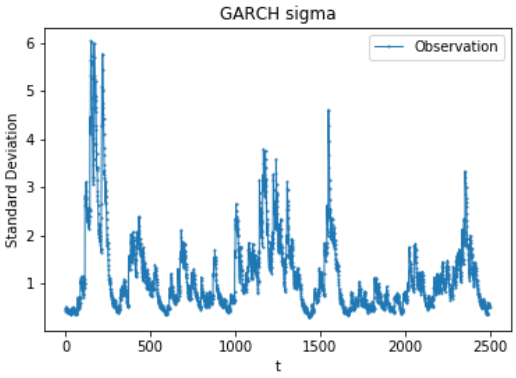

Figure 4: Conditional standard deviation σ_t in the simulation analysis of Bayer and Dimitriadis (2018)

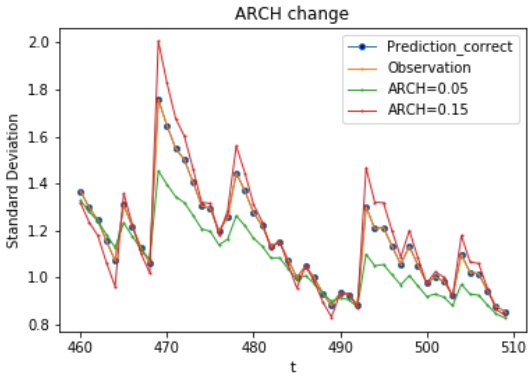

Figure 5: Conditional standard deviation values under correct specification (orange & blue dots) and several misspecifications of the ARCH parameter

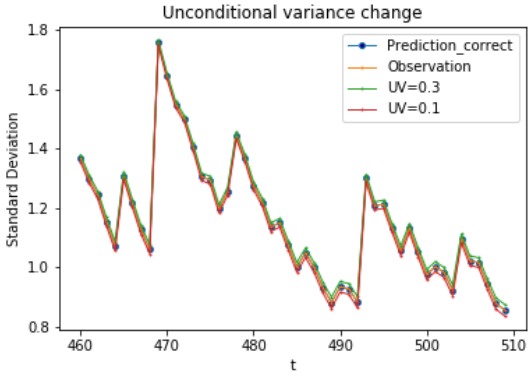

Figure 6: Conditional standard deviation values under correct specification (orange & blue dots) and several misspecifications of the unconditional variance

(2018).

The misspecification of the ARCH parameter makes the model less accurate (see Figure 5), but it does not cause structural underestimation of the variance. It changes the impact of the returns on

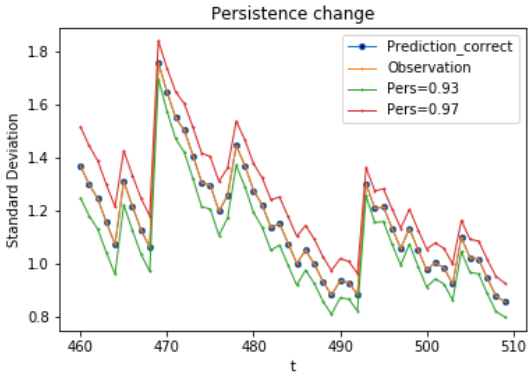

Figure 7: Conditional standard deviation values under correct specification (orange & blue dots) and several misspecifications of the persistence

the conditional variance, which mostly makes the model 'overshoot' the true values, due to the large jumps upwards compared to the many smaller drops downwards, visible in Figure 5. Because of this, an increase in the ARCH parameter usually leads to an increase in the estimated conditional variance, with a decrease in ARCH parameter acting in the opposite direction. This would suggest that lowering the ARCH parameter decreases the acceptance rate of the backtesting methods, while increasing it can increase the acceptance rate (which contradicts Bayer and Dimitriadis, who state the opposite).

If we consider the second scenario (Figure 6), we expect to agree with Bayer and Dimitriadis' analysis that a decrease in unconditional variance will decrease the acceptance rates of the backtesting methods. This is because this change very consistently makes for underestimation of the conditional variance, and thus, underestimation of tail risk. An increase of the unconditional variance, of course, has the exact opposite effect.

If we look at the last scenario, change in shock persistence (Figure 7), we see that a higher level of persistence gives overestimation of conditional variance overall, while a lower level gives underestimation. This is, as in scenario (1), opposite what Bayer and Dimitriadis state. Thus, we expect the backtesting methods to give lower acceptance rates when the level of persistence is lowered.

4.3.2 Impact on acceptance rates

Here, we present the size-adjusted power graphs. Again, we must note the inconsistency in jumps on the x-axis in all figures. Due to this, the curvatures of the methods can be misleading. As in Section 4.2, we care more about the relative differences in performance, rather than the individual curvatures.

We also note on the LWZ method that it breaks for large sample sizes. When $T = 252$, what we tend to use in most of the analyses, this is not a problem. For this particular section, though, we use $T = 2500$, which makes the setup very numerically unstable. Thus, we must cap the simulation length to $T^* = 500$ for the LWZ part of the analyses. This does make the results for this particular method less reliable than for the remaining backtests. In practice, however, one tends to deal with one or two years

of data, wich equals 252 to 504 datapoints, thus rendering this issue irrelevant for general backtesting purposes.

From Figure 8 we see that the GP and MP methods have the best performance, since their size-adjusted power goes up the most due to a decrease in the ARCH parameter $\tilde{\gamma}_1$. In the most extreme misspecification, their power even reaches the convergence point of 100%. Furthermore, we see that the DMP, RC and BD* methods have very comparable size-adjusted power levels, but are slower to react to the misspecification than the two mentioned above. Finally, the LWZ method seems to split the difference, coming out in the middle of the group. When the ARCH parameter increases instead, all methods' rejection rates drop below the 5% level, as expected; this effect is most noticable for the GP, LWZ and MP methods. When the parameter increases too much, though, the rejection rates curve back up to around 5-6%.

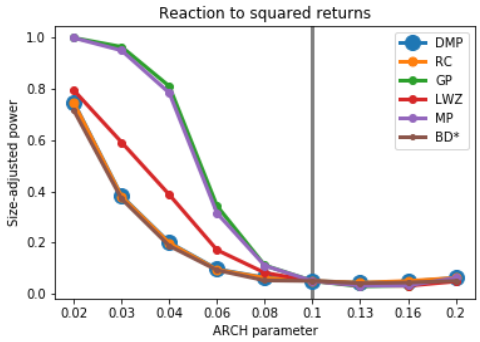

Figure 8: Rejection rate (power) per backtesting model plotted against increase in reaction to squared returns, grey vertical line represents true model; B&D setup

From Figure 9 we see again that the GP and MP methods are the most powerful, since they are the most sensitive to changes in the unconditional variance. This time, GP slightly but visibly outclasses MP in the more extreme misspecification scenarios. Again, LWZ performs better than the remaining three methods, but this time BD* performs significantly worse than RC and DMP, seemingly having hardly any sensitivity to this type of misspecification. Contrary to Figure 8, the methods do not seem to converge on a power level of 100% under the extreme misspecification setting. The methods all do seem to converge on very low rejection rates in the variance overestimation setting though, which is to be expected in a one-sided testing setup.

Lastly, Figure 10 shows the same behaviour as before for most of the methods; GP and MP are much more powerful than the other methods, with hardly any difference between the two. LWZ is the third most powerful, and RC, DMP and BD* have approximately the same performance. These last three methods' sensitivity seems to pick up in the extreme scenario, as opposed to the flattening off of the power level in Figure 9.

Again, the methods seem to converge on acceptance rates near 100% in the risk-overestimation scenario, which is expected and desirable.

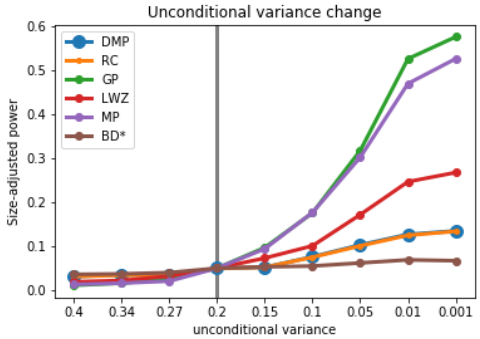

Figure 9: Rejection rate (power) per backtesting model plotted against decrease in unconditional variance, grey vertical line represents true model; B&D setup

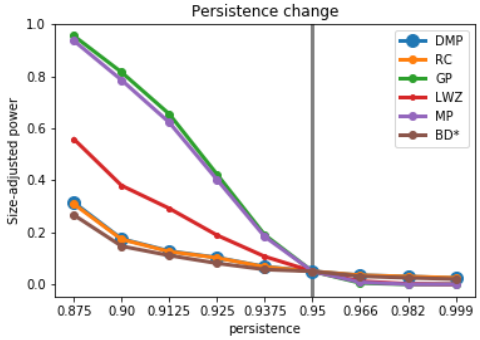

Figure 10: Rejection rate (power) per backtesting model plotted against increase in persistence in shocks, grey vertical line represents true model; B&D setup

Overall, the results of the size-adjusted power analyses consistently indicate that the GP and MP methods are the most powerful, with LWZ being third and the RC, DMP and BD* methods lagging behind significantly.

We also observe sensitivity in the wrong direction (compared to Bayer and Dimitriadis' analysis) to changes in two of the misspecification categories. This is due to the fact that an increase in reaction to squared returns does not necessarily cause underestimation of the tail risk, and an increase in persistence causes overestimation of tail risk instead of underestimation.

This setup differs compared to the previous analyses in that it does not take 'clean' t-distributed residuals, but the Y-variable observations of the GARCH model as input for the backtests. Furthermore, this setup uses a time period of 10 years (T = 2500 observations), which is much larger than the sample size for our previous analyses. Lastly, the conditional variance is very volatile under this specification, as can be seen in Figure 4. Although the predicted values $\hat{\sigma}_t$ follow the true conditional standard deviations σ_t rather well under most of the misspecification scenarios, we do observe structural biases and inaccurate sensitivities to shocks in the graph. This is of course exactly what is intended for this analysis, and our backtesting methods should be able to notice these misspecifications.

4.4 Variance-Covariance simulation

We present the results of our final analyses in this section, reporting on the findings of the Variance-Covariance setup. The first two figures, Figure 11 and 12 are rather straightforward. We see that an increase in variance of the portfolio through an increase in either the correlation coefficient ρ or the standard deviation parameter σ_2 triggers the backtesting methods to detect underestimation in tail risk. Again, we see the two methods of MP and GP clearly perform very well, with the methods of DMP, RC and BD* underperforming significantly in terms of size-adjusted power. What's different in these analyses compared to those of Section 4.3, though, is the fact that the LWZ method has the same power level as MP and GP. Furthermore, we see that the power of the VaR backtest is slightly below the top three contenders of MP, GP and LWZ.

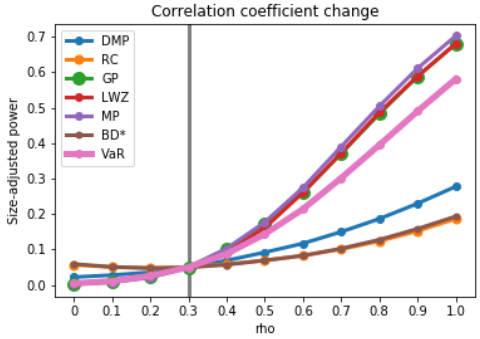

Figure 11: Rejection rate (power) per backtesting model plotted against increase in correlation coefficient ρ, grey vertical line represents true model; Variance-Covariance setup

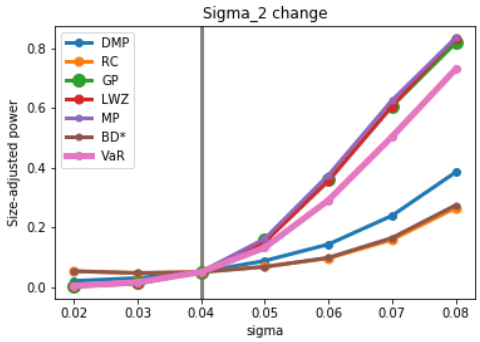

Figure 12: Rejection rate (power) per backtesting model plotted against increase in standard deviation σ_2, grey vertical line represents true model; Variance-Covariance setup

When considering Figure 13, we see more interesting curvatures. The ordering of methods' powers is still the same, but there is less difference between the VaR backtest and the MP, GP and LWZ methods than before and there seems to be more of a gap between these and the group of DMP, RC and BD*. Lastly, all methods but RC and BD* seem to react as expected to the upwards shift in mean, since

this makes for overestimation of tail risk. Only these last two methods follow almost exactly the same concave curvature for a slight misspecification of the mean in the case of risk overestimation.

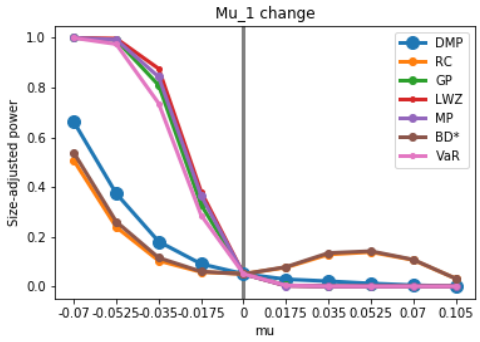

Figure 13: Rejection rate (power) per backtesting model plotted against increase in mean μ_1, grey vertical line represents true model; Variance-Covariance setup

We note the following regarding the comparison of our ES backtesting methods with the VaR framework. The GP, MP and LWZ methods are all slightly more sensitive to misspecifications in this bivariate normal setup than the VaR backtest that utilised the same underlying prediction mechanism, namely the Variance-Covariande method. If we extrapolate, this observation gives an indication that applying any of these methods when the risk measure reporting change is fully adopted, will result in a slightly stricter regime in terms of tail risk model evaluation than is currently in use. Of course, this is hardly conclusive evidence to this hypothesis, but more research into this area could prove interesting.

4.5 Implementational complexity

In this section, we report on the ease at which the methods can be implemented. This entails the difficulty in terms of programming and understanding of the method, as well as computation time and input requirements of the algorithms.

4.5.1 Computation time

Not all methods that we analyse have the same straightforward structure as, for instance, a t-test would. The methods by Righi & Ceretta and Moldenhauer & Pitera compute a threshold value as a separate a priori step, by simulation under the hypothesised distribution that the ES prediction was based on (using MC methods). This is comparable to finding the appropriate rejection threshold for counting exceedances in the VaR framework. This value is then compared to the statistic computed from observed data to decide on acceptance or rejection. The determination of the threshold value generally demands much computation time, but the calculation of the statistic for the observed data is usually near instant. Because of this, the computation time required for a single evaluation of an ES prediction model will be relatively large for the RC and MP methods, compared to the other methods. However, if evaluations must be done regularly, say, every day, on the same prediction model, then these evaluations can be

done much faster for the RC and MP method compared to the others. From this, it naturally follows that the RC and MP methods easily win out in terms of computation time in the long run. The caveat here is that this only goes so long as the prediction model remains identical; if any of the parameters are adjusted or the underlying distribution changes, the initial calibration step of the backtesting method has to be performed again.

The BD* method also requires MC iteration, suggesting that its computation time will be of a larger order of magnitude than the remaining methods. Finally, some methods might be computationally complex to such an extent that they take a relatively large amount of time to perform, which is what we see in the LWZ method. We discuss the cause of this more extensively in Section 5.4.

Now for the results; the computation time can be quantified with relative ease, of course, as is done in Table 3. Note that the RC and MP methods spend a relatively large amount of time on finding the appropriate rejection threshold values for the specified null distribution (in brackets in Table 3), but checking the observed test statistic against this critical value costs very little time for these methods, as discussed above. Since the threshold value calculation is only required once per assumed underlying distribution (only t_5 is used), and not for every separate null hypothesis setting, this computation time is not of much consequence for this test setup. The computation time of the observed test statistic, on the other hand, has to be repeated for each scenario, as it takes both the observed values and the ES predictions as input. When performing this method in a real scenario it is probably fair to take the combined computation time of these parts, though. Furthermore, the LWZ method only uses one fifth of the observations per iteration (due to computational limitations), which suggests that the computation time should be multiplied by five for an appropriate comparison to the other methods.

We see that the methods by LWZ and BD* have computation durations far beyond those of the other methods, though we noted above that the fair LWZ computation time is much greater still. The RC and MP methods' critical threshold calculations are rather similar in duration when the same amount of iteration simulations is used (10^5 in this case). The computation time of GP is rather reasonable; and finally, DMP's speed is on par with the test statistic part of the RC and MP calculations, which is negligible in comparison to all the other computation durations.

Method	DMP	RC	GP	LWZ	MP	BD*
Computation time	0.82	0.12 (44.94)	6.04	149.39	0.37 (64.49)	191.33
Variance	0.03	0.0 (27.54)	0.06	68.15	0.01 (62.99)	295.39

Table 3: Computation time (in seconds) and variance per backtesting model for one scenario of the B&D simulation setup ($N = 1000$, $T = 2500$); values in brackets represent computation time for the rejection threshold under the null hypothesis, where applicable

The complexity of the methodologies cannot be quantified in such a straightforward manner as the other criteria to rate a backtesting methodology on. However, we can discuss the methods as described in Section 3.1 rather well in a subjective manner. The DMP method is clearly rather trivial, in that

it hardly differs from the commonly used t-test for location of mean. It uses a slight adjustment for the distance between the sample mean and the ES estimation and the variance and dof used for the calculation of $\hat{t}$ are only based on the exceedances, instead of the whole dataset. The RC method also employs what seems to be derived from the t-test, by taking the exceedance residuals and testing for equality of their mean to the predicted ES value, dividing by the assumed standard deviation of the exceedance under the null hypothesis. The added value here, compared to the DMP method, is an MC simulation for finding the appropriate rejection threshold.

The MP method is tough to compare to the others, because of its drastically different approach. The method is very intuitive in its way of evaluating the validity of the ES prediction, through its simplicity and elegance. It also makes use of an MC simulation for the rejection threshold, which is why it ranks higher than than DMP in terms of complexity, but it is on par with the RC method in this regard.

The final methods of BD*, GP and LWZ are all very complex in their own regard, and they are difficult to compare against each other because of it. In our experience, the LWZ method took the most time and effort to implement correctly, but none of these methods are practical for widespread use outside of well-polished and highly tailored software packages.

Lastly, we note on the required input for the methods. Apart from the obvious necessity of all method to receive the observed data as input, the required inputs are given in Table 4. Interestingly enough, the LWZ method is the only one that does not require the ES predictions, but rather test the validity of the entire tail distribution. We also see that GP requires a lot of information, whereas MP and BD* require rather very little.

Required input	$\hat{ES}$	$\hat{VaR}$	σ^2 (tail)	μ	distribution
DMP	YES	YES	NO	YES	NO
RC	YES	YES	NO	NO	YES
GP	YES	YES	YES	NO	YES
LWZ	NO	NO	YES	YES	YES
MP	YES	NO	NO	NO	YES
BD*	YES	YES	NO	NO	NO

Table 4: Input requirements per backtesting method, in addition to the observed data

5 Discussion

In this section we discuss caveats and limitations that we uncovered during the research. We also make recommendations for future research on the subject as a whole and specifically on some of the backtesting methods that, in our opinion, could be improved in certain aspects.

One suggestion for future research, that we considered but did not get to due to time constraints,

is a traffic light system for the ES framework as is also in place for the VaR framework. This can be based on the p-value, using $\alpha_2 = 10\%$, for instance as the starting point for a danger zone; there are already papers that suggest applying this to their suggested ES backtesting method. This can also be applied to the MP method in a very similar manner to the way it is done in the VaR framework, since the MP method also uses a discrete test statistic. Therefore, a discrete range can be determined as a danger zone, just as in the current VaR framework.

5.1 Empirical analyses

One remaining point of discussion is the use of historical simulation. Here, we discuss why using real data would be relevant and beneficial. One of the reasons is that the performance of a backtest depends heavily on what prediction model and what underlying data it concerns, as can be concluded from Garcia-Jorcano (2017). To get the most accurate backtest evaluation, the analysis would thus need to be performed on real market data. Another important factor is that regulators will not be performing backtests on financial institutions' ES estimates themselves, but simply periodically check up on the backtesting methods and results that these institutions produce, in order to determine minimum capital requirements for these institutions (as gathered from statements by professionals at the risk management department of EY Amsterdam). Banks have been allowed to use internal models as a basis for calculating their market risk capital requirements since 1997 (BCBS, 2014). These institutions will very likely also be free to estimate their ES in whichever way they see fit in the new framework, but the backtesting method must be approved by the regulator before use. From this we can conclude that we must be able to backtest any ES prediction model, regardless of the underlying (distributional) assumptions used. Therefore, performing an analysis on historical data rather than using parametric simulation would be very insightful, if performed correctly.

Unfortunately, as of yet we do not have a procedure available for performing an empirical analysis in such a way that we can control for the accuracy of the historical data-based ES predictions. We have debated extensively on this, but we ultimately decided not to tackle this problem. We do have a suggestion for dealing with the issue in future research, making use of the VaR backtest. If we assume that the VaR backtest is sound, then we only need to investigate whether our to be analysed ES backtest rejects predictions when the VaR backtest also does so on the VaR prediction using the same underlying mechanics as the ES prediction. Still, this means we can only really analyse one backtest when assuming another is always completely valid, but this seems much more valid when the 'trusted' backtest is the one which we have been using for the past twenty odd years, with tons of empirical evidence to its name, rather than any other handpicked ES backtest that has yet to prove its worth in financial regulation. On the other hand, this rather undermines the whole point of moving away from VaR backtesting and towards ES backtesting, if we calibrate all our ES backtests on the original VaR backtests.

5.2 Moldenhauer and Pitera

Although the methodology of Moldenhauer and Pitera is quite straightforward and elegant, there is one caveat to it. This comes in the form of what this method tests for, because it takes the set of exceedances, and then tests how many of the observations just above the VaR threshold have to be added to this set for the combined mean value to be larger than the mean of the ES predictions for all those observations. It is not a drawback, necessarily, but it makes for a situation where the acceptance or rejection of the ES of the lowest ν quantile is dependent on the observations just above this quantile. Thus, there is an implicit assumption in play that more than just the outer 2.5% of the tail losses consists of relevant observations, the shape of which concerns us. This also suggests that we implicitly assume that more than just this 2.5% is one continuous distribution as specified under the null hypothesis, whereas most methodologies only strictly test the outer 2.5% for adherence to this hypothesised distribution.

5.3 Bayer and Dimitriadis

We recommend future research on this topic to go deeper into the issues of the Bayer and Dimitriadis regression backtest. As noted earlier, the bootstrapped sample variance used in this methodology has serious issues in terms of validity. Research into the comparison between this method and our recommended alternative of the original sample variance could be invaluable to the literature on this subject, along with general research on the stability of bootstrap sample variances.

5.4 Löser, Wied and Ziggel

Despite the fact that the LWZ method does not employ any MC iteration, its computation time has been disproportionately long throughout the simulations. This is due to the complexity of the algorithm, which performs a transformation of the data to a uniformly distributed set of binomial draws of exceedances of the VaR threshold (under the null hypothesis), employing the Irwin-Hall distribution (Hall, 1927). The cumulative probability of the observed data is calculated as the sum of the probability mass function (pmf) of the binomial distribution multiplied by the uniform transformation of the binomial successes as outer left tail observations for all possible amounts of binomial successes that could have occurred. For $T = 252$, this is a sum of $k = 1$ to 252 of the binomial pmf: $\binom{T}{k}\nu^k(1-\nu)^{T-k}$ multiplied by the CDF of the Irwin-Hall distribution. This is so numerically unstable that it is very difficult to evaluate. However, this method would be very powerful if this computational issue can be avoided through an elegant approximation of this huge sum of Irwin-Hall multiplied by binomial values. Therefore, I highly recommend future research to be aimed at cracking this intricate combination of distributions for a powerful and versatile ES backtest.

5.5 Variance-Covariance analysis

The two-stock portfolio simulation that is done presents a simple and clear analysis on the currently widely used Variance-Covariance method for VaR and ES estimation and its backtesting. However, one could note that the two-stock portfolio of normal distributions simply provides a normally distributed variable r_t. Although the (assumed) parameter specifications of the two underlying stocks are in fact used in the Variance-Covariance estimation method, the resulting estimates could easily be obtained through the use of the one random variable too. Since it is probably also possible to reconstruct the distortions that we applied in this analysis through the use of a single normal distribution for r_t, it would be insightful to apply more extreme distortions to just one of the stocks in the portfolio, such that the resulting scenario cannot simply be reconstructed through the use of the single random variable (e.g. a change from one of the stocks normal dgp to a t-distributed process). This would allow us to observe a stronger effect of partial misspecifications in the portfolio on the Variance-Covariance estimates of VaR and ES and the backtest evaluations.

6 Conclusion

This research is on the topic of the change in risk measure from VaR to ES. The ES initially lacked a proper backtesting method, but many papers have been published since the early 2000's describing methodologies for backtesting this statistically difficult risk measure. The main problem here is the small sample size of extreme tail events, that this risk measure inherently deals in. We analyse a set of six of these methods, choosing those that are the furthest developed along their niche approach to the backtesting issue at hand.

The core of the research is the analyses of these methods in a variety of simulation settings. These are mostly focused on normal or t-distributed random variables, with slight but deliberate mismatches in parameters such as variance and dof between the hypothesis underlying the predictions and the dgp. All of the analyses are done in a one-sided testing setup, since we only care about detecting underestimation of tail risk, whereas we are not interested in its potential overestimation. Through this, we gain insight into the relative performances of all of these methods, using size-adjusted power as our main performance measure. Secondary measures for adequacy of the methods are computation time and ease of implementation, given that we are looking for backtesting methods that can be implemented throughout the Risk Management sector of the financial world.

We observe that the methods by Graham and Pál and Moldenhauer and Pitera clearly outperform the others throughout the simulation analyses in terms of size-adjusted power. A close runner-up is the Löser, Wied and Ziggel method, though this method has issues when applied a sample of $T > 500$. It performs very adequately and comparably to the GP and MP methods in normally distributed setups, but drops off in the Bayer and Dimitriadis simulation setup 4.3 which has a more complex underlying distribution, using a GARCH framework.

When considering the difference between the GP and MP methods, we conclude that the MP method reigns supreme. This is due to the fact that it is much more easily implemented through its relative mathematical simplicity. Furthermore, it is the faster method of the two in terms of computation time in the long run, for daily evaluations on a static prediction model throughout an entire year. Lastly, the GP method showed some inconsistencies in situations of extremely erratic behaviour of conditional variance and error term in the GARCH setup.

Our last analyses show that the recommended MP backtest for ES is slightly, but consistently more sensitive to misspecifications of the underlying distribution than the basic VaR backtest is. This comparison is not completely straightforward, but the $ES_{0.025}$ and $VaR_{0.01}$ predictions can be made based on the same mechanics and assumptions, and are almost identical under normality. Therefore, misspecifying the underlying mechanics and/or assumptions distorts both risk measures in exactly the same way, and thus, the comparison between the sensitivity of the ES and VaR backtests can be made.

The results of these analyses indicate that the regulatory risk management standards will be slightly stricter under the ES framework than they were under the VaR framework, if the MP method will be employed. This could make for higher and more frequent fines for financial institutions, since the regulators will be upholding a stricter risk management regime.

Lastly we note a limitation to our research. We are as of yet unable to find a way of properly evaluating backtesting methods' validity when using empirical observations. Thus, all of this research is based on simulated scenarios where the data follows known distributions. Therefore, we cannot say for certain if the methods discussed in this paper are appropriate in reality, without having to make assumptions on the underlying distributions.

A Appendix

A.1 Size-adjusted power

For a fair comparison of power of backtesting methods, an adjustment for size can be made (Lloyd, 2005). This is done by adjusting the critical value of the test statistic corresponding to the backtest in such a way that the null hypothesis is rejected exactly $\alpha = 5\%$ of the time when the test is performed on simulated data drawn from the null hypothesis. Ideally, this is done through an invertible function $G(t)$, but with some trial and error of adjusting the critical value of the backtest, the appropriate percentage of rejections can be obtained too. The only instances where this is impossible (for our methods) are the discrete rejection thresholds of the method by Moldenhauer and Pitera (2018) and the VaR backtesting method. We can illustrate this with relative ease by the example of the Bayer & Dimitriadis simulation analyses 3.2.3. In this setup which utilises draws from the t_5-distribution with $T = 2500$ observations, the rejection threshold value for the MP method equals 83, giving a rejection rate (or size) of 4.5% when the null hypothesis is specified correctly. For the appropriate size-adjusted power analysis we need a rejection rate of 5%, but if we lower our critical value to 82, we get a rejection rate of 5.3% (anywhere between 82 and 83 gives an equivalent result to using 83 in this discrete setup). To solve this issue, we use a weighted average of the scenarios of threshold values 82 and 83 to obtain the exact size of 5%, which will in turn give us the size-adjusted power that we are looking for in the scenarios of misspecification.

Note that, strictly speaking, the applied methodology for obtaining size-adjusted power levels in the Bayer & Dimitriadis simulation analyses is not the same as what Lloyd (2005) describes. According to Lloyd, the size adjustment of the real data must be calibrated on simulated data under the null hypothesis. This suggests that the null hypothesis be kept constant, while the data input changes. What we do, however, is change the assumed specification of the conditional variance and thus the null hypothesis, while the data stays the same. This adaptation might give slightly different size-adjusted power levels from what the methodology by Lloyd (2005) would suggest. It is much less time consuming to implement, however, since the size calibration is only required once per method instead of separately for each method for each of the 27 null hypotheses. In analysis 3.2.4 we keep the null hypothesis constant, while changing the input data, thus following Lloyd's methodology.

A.2 Acceptance rate tables

The following are the full tables of results from the preliminary analysis on all six backtesting methods. The bottom left values are expected to all be (very close to) 1, since we are running one-sided backtests and these scenarios correspond to the overestimation of the tail risk, which we are not interested in. These are reported as a check on whether the confidence bounds/rejection regions are specified correctly.

assumed \true	Normal	t_3
Normal	0.9589	0.0225
t_3	1.0	0.9765

Table 5: Acceptance rates Righi & Ceretta; for $N(0,1)$ and t_3 tail scenarios

assumed \true	Normal	t_3
Normal	0.9531	0.0223
t_3	1.0	0.9531

Table 6: Acceptance rates Graham & Pál; for $N(0,1)$ and t_3 tail scenarios

assumed \true	Normal	t_3
Normal	0.9619	0.2066
t_3	1.0	0.9809

Table 7: Acceptance rates Del Brio, Mora-Valencia & Perote; for $N(0,1)$ and t_3 tail scenarios

assumed \true	Normal	t_3
Normal	0.8878	0.5258
t_3	1.0	0.9998

Table 8: Acceptance rates Löser, Wied & Ziggel; for $N(0,1)$ and t_3 tail scenarios

assumed \true	Normal	t_3
Normal	0.9570	0.0675
t_3	1.0	0.9617

Table 9: Acceptance rates Moldenhauer & Pitera; for $N(0,1)$ and t_3 tail scenarios

assumed \true	Normal	t_3
Normal	0.990	0.610
t_3	1.0	0.999

Table 10: Acceptance rates Bayer & Dimitriadis; for $N(0,1)$ and t_3 tail scenarios

θ	∞	30	22	15	10	7	5	3
BD	0.935	0.8305	0.769	0.6425	0.446	0.286	0.2785	0.905
BD*	0.942	0.823	0.756	0.5885	0.3355	0.1225	0.0155	0.0

Table 11: Acceptance rates Bayer & Dimitriadis original (BD) and adjusted (BD*); Löser et al. setup

A.3 Bootstrap confidence interval

One of the most well-known methods of improving the usefulness of limiting amounts of data is the bootstrap method (Efron & Tibshirani, 1994). If we use this to resample our data in a random way, we can get a 95% confidence interval (CI) of the observed shortfall to compare with an ES estimate. If the estimated ES is less extreme than the mildest 5^{th} percentile of these resampled shortfalls (for a one-sided test), then the bootstrap method suggests rejection of the ES estimate. This is one of the simplest ways of testing the ES estimate, since it only depends on the observed data of the year to be tested for. The bootstrap method just randomly draws from this observation pool to find our rejection region, and no supporting assumptions are needed. One obvious downside is the computational time, which is rather large due to the iterative resampling that this method employs.

First, we define a set of P&L observations $O = \{o_1, ..., o_n\}$, with $n = 252$ trading days in a year. The VaR exceedances of the observed P&L's are denoted as E:

$$E = \{o_i \in O : \ o_i < q_i(\nu)\}_{i=1}^{n}. \tag{44}$$

Note that we suspect an issue with bootstrapping rare tail loss events, causing major inaccuracies in variance approximation. This is beyond the scope of the research, and thus has not been investigated further, though there might exist literature on this topic already.

A.3.1 Bootstrap results

The bootstrap method is much slower in the Wimmerstedt setup than the other methods that we analyse. It takes a large amount of resamplings of the randomly generated (standard normal or t-distributed) data for the bootstrap method to get accurate rejection regions ($B = 10^3$ resamplings used); and the MC simulation, on top of that, also requires a large number of iterations of random draws of data ($N = 10^4$ iterations used) in order to converge to a stable acceptance and rejection rate. In the other methods that are discussed in this section, only the MC iteration amount N is relevant, but this effect compounds for the bootstrap method.

Interestingly, using $B = 10^2$ and $N = 10^3$ takes about the same computational time as using $B = 10^4$ and $N = 10$. This suggests that computational time scales by the product of the iteration amounts, or $T \propto B * N$. This seems intuitive, since 10 times 10000 bootstraps and 1000 times 100 bootstraps equals the same amount of bootstraps. This proves to be approximately true when taking a factor 10 or 100 more iterations for either B or N, resulting in 500 and 5000 seconds of computation time, respectively (some algorithm optimisation for speed improvement can be applied here, though). It also seems that using a low amount of MC simulations has much more influence in the significance of the acceptance and rejection rates, than using low bootstrap amounts. Theoretically, the amount of bootstrap iterations to run only influences how well one approximates the confidence level $\alpha = 5\%$, with $B = 10$ giving a very crude rejection if $H_0 \leq 9$ out of 10 sample averages, and $B = 10^2$ already giving a rather decent rejection if $H_0 \leq 95$ out of 100 sample averages. The amount of MC iterations is more

important for the accuracy and significance of the rejection rate that the given confidence level grants us; for instance, when performing the method several times with $N = 10^3$, the second decimal of the resulting acceptance and rejection rates fluctuates greatly between runs. We suspect this is an issue of accuracy versus consistency. Thus, $10^2 \leq B \leq 10^3,\ 10^4 \leq N \leq 10^5$ is used, in the interest of time.

The results of the simulation are given in table 12. Getting these values took more than 80 minutes (with $B * N = 10^7$), and these values still fluctuate more on repeated runs of the script than those of the other methods.

Bootstrap	Acceptance when true	Rejection when false
$B = 10^3$	0.8824	0.9153
$B = 10^2$	0.8785	0.9210

Table 12: Acceptance/rejection rates of the bootstrap method (using $B * N = 10^7$); for the false scenario $t_3(0, 1)$ is used

For such a simple method, these rates are not all that bad. To put these in perspective, the 'correct' Emmer method, as shown in table 14, has a slightly higher power of accepting true scenarios, but is slightly worse at rejecting false scenarios. Thus, despite the computational time issue, this seems like a good benchmark to try to beat.

What is really interesting about this method, though, is that bootstrapping can be done in much more elaborate ways, which will be especially useful when dealing with portfolios of data. Thus, this might offer a reasonable, 'simple' alternative to tail simulation, which is very complex for portfolios of intricate derivatives.

A.4 Bayer and Dimitriadis: two-sided Backtest

Here, we discuss the two-sided regression based backtest by Bayer and Dimitriadis. The Wald statistic is used:

$$T_{ESR} = \Big((\hat{\beta}_0^e, \hat{\beta}_1^e) - (0, 1)\Big)\hat{\Sigma}_{ES}^{-1}\Big((\hat{\beta}_0^e, \hat{\beta}_1^e) - (0, 1)\Big)', \tag{45}$$

This test statistic (asymptotically) follows a χ^2 distribution with $\theta = 2$:

$$T_{ESR} \xrightarrow{d} \chi_2^2. \tag{46}$$

The estimation of β_0 is done by optimisation of the loss function ρ:

$$\rho(Y_t, X_t, \beta) = \frac{1}{-X_t'\beta^e}\left(X_t'\beta^e - X_t'\beta^q + \frac{(X_t'\beta^q - Y_t)\mathbb{I}_{\{Y_t \leq X_t'\beta^q\}}}{\tau}\right) + \log(-X_t'\beta^e), \tag{47}$$

where $X_t'\beta^q = \beta_0^q + \beta_1^q\hat{e}_t(\nu)$, and $X_t'\beta^e = \beta_0^e + \hat{e}_t(\nu)$. The following optimisation is done with respect to β^q and β_0^e:

$$\widehat{\beta_T} = argmin(\beta)\ \frac{1}{T}\sum_{t=1}^{T}\rho(Y_t, X_t, \beta). \tag{48}$$

We use the Powell method in the *scipy.optimize* package, for simplicity. Perhaps it is worth looking into faster methods eventually, if this algorithm takes too long.

A.5 Relevance

This must still be explained: [There are plenty of ES backtesting procedures already developed; why is finding a backtest still a problem?] - methods are not effective enough (in terms of true acceptance and false rejection, type 1 and 2 errors) and/or they have unrealistic assumptions (mostly on the shape of the tail distribution, which does not follow any known type of distribution [citation needed]). Furthermore, the literature on backtesting ES is fairly new, thereby leaving a potential loophole for errors in ES computations to go unnoticed for a while because financial industry applications may not yet perform routine and powerful enough ES backtesting (Acerbi & Szekely, 2014), (Nesmith, Oh et al., 2017).

This research is relevant primarily for practical applications. The Basel guidelines are moving from VaR as a regulatory tool towards ES, for a number of reasons. The main reason for this change having been rather slow over the past decades is that ES is a much more difficult risk measure to backtest than VaR. This research aims to adapt current methods of backtesting by solving the problems accompanied with these methods that were not dealt with adequately in previous literature. This would be of much interest to any regulatory authority that deals with market risk. Besides, it is of much use to any organisation that deals with market risk and wishes to check the accuracy of their models for measuring this risk.

I will focus on keeping these methods as generally applicable as possible, since that will be the most useful in a general framework. In contrast, having one very powerful backtest for an ES prediction based on a very specific underlying distributional assumption would be useful for internal backtesting at whichever organisation makes use of this distribution for their tail loss predictions, but will be rather useless for any other ES prediction scenario. This is because a backtest in such a scenario can be done in a relatively straightforward manner, by assessing the probability that the observed exceedances were drawn from the distribution that was used for construction of the ES prediction. This would be a simple goodness-of-fit test of the data on the assumed distribution, and because of this, would only be valid for backtesting in the case that this exact distribution is used for estimating tail loss. Thus, one would be backtesting a single ES prediction methodology, which is completely irrelevant for all other ES prediction methodologies.

Depending on what methods can be found to perform these backtests of ES well, the research could be very interesting from a scientific point of view, too. If only a rather sloppy, ad-hoc method of backtesting in a general framework can be found that lacks any sound theoretical background, this will not have many scientific implications. If, however, we manage to find sound statistical grounds for a significance test of some sort, this will most likely have consequences for the theoretical way we look at tail distributions of P&L data, and by extension, the theory behind shocks spot prices of stocks.

To conclude, there is much interest in this field from both a theoretical and a practical standpoint,

but the relevance of this research will depend, of course, on the type of findings.

A.6 Extension on Emmer, Kratz and Tasche

One idea, that is seen in Emmer et al., is looking at how VaR backtests work. These simply count the VaR-threshold exceedances, which follow a binomial distribution with a success probability equal to the VaR-percentile, by construction. We then simply compare the observed proportion of exceedances to its expected value, and test whether this difference is significant. This VaR test has the luxury of relying on the entire sample in this way, while ES backtests can generally only use the extreme loss tail as useful sample space.

The main concept in Emmer et al. is to split the ES into a number of separate VaR-percentiles, which can be interpreted as a crude approximation of a discretization of an integral, where the ES is (theoretically) the integral of the tail loss of the p&l distribution, ranging from minus infinity to the VaR-threshold. All that you are left with then, is a number of VaR-percentiles to backtest, which can be done rather easily with the aforementioned binomial approach.

We can use several methods of combining the information of all these binomial distributions to produce one single p-value (Heard & Rubin-Delanchy, 2017); Emmer's method of rejecting the nullhypothesis if any of the p-values is lower than α is basically a rough version of Tippett (1931)'s method: $S_T = \min(p_1, ..., p_m)$. Perhaps it is not as sophisticated as Tippett's method, though, since it does not use the Beta rejection region that they suggest this measure follows ($S_T \sim \text{Beta}(1, m)$). Another way of combining these p-values is using Fisher's combined probability test, which uses the following statistic:

$$-2\sum_{i=1}^{m} \ln(p_i) \sim \chi^2_{2m}, \quad (Fisher,1934) \tag{49}$$

where p_i represents the p-value of the i-th VaR-percentile. We use $m = 5$ percentiles, starting at the 2.5-th percentile and moving down in increments of 0.5. What this gives is a single measure that follows a Chi-squared distribution with dof $2m = 10$.

As a sidenote, more sophisticated (and newer) methods are also available, like that of George and Mudholkar (1979): $S_G = S_F + S_P = \sum_{i=1}^{m} \log\{p_i/(1-p_i)\}$.

A concern that arose when using Emmer's method is that the VaR-exceedances are rather dependent on each other (Wimmerstedt, 2015). Starting from the largest VaR-percentile of 2.5, all subsequent percentiles, e.g. 2, 1.5 ..., are fully contained within the previous one, and thus their exceedances are dependent on those of the previous level. The method by Fisher, however, relies on the p-values to be independent. This issue can be solved in two ways, the first being to condition the binomial distribution of the $2.5 - 0.5i$-th percentile on the amount of exceedances of the previous percentile. The second is to define intervals that do not overlap. We define X_i as the observations in the following interval:

$$x_i \ \in \ [VaR_{0.025-0.005(i-1)}, \quad VaR_{0.025-0.005i}]. \tag{50}$$

Using this, we wish to test whether all of these disjoint regions contain the appropriate amount of observations. Under the nullhypothesis that our VaR-percentiles are correct, all of these intervals should

contain 0.5% of our data, or $\forall\ i \in \{1, ..., 5\}:\ X_i = X = 0.005n$, with n being equal to the sample size. Furthermore, the proportion of successes[6] follows, as previously mentioned, the binomial distribution, with $p = 0.005$. For independent binomial distributions, we can use the following cumulative probability function:

$$P(X_i \geq x_i) = 1 - F(x_i - 1; n, p) = 1 - \sum_{k=0}^{x_i - 1} \binom{n}{k} p^j (1-p)^{n-k}; \tag{51}$$

however, the disjoint intervals are still correlated, thus we must condition x_i on $x_1, ..., x_{i-1}$. This gives us:

$$p_i | I_{i-1} = p(X_i \geq x_i \mid x_1, ..., x_{i-1}) = 1 - F\Big(x_i - 1;\ n_i^*,\ p_i^*\Big); \tag{52}$$

where I_i represents the information of $x_1, ..., x_i$. This gives us adjusted values for our remaining sample size n^* and proportion p^* given I_{i-1}:

$$n_i^* = n - \sum_{j=1}^{i} x_j, \quad p_i^* = \frac{p}{1 - (i-1)p}, \tag{53}$$

for the calculation of independent p-values to be used for a combined probability method. If the p-values are not independent, however, Brown's method can be used as an extension on Fisher's method (Brown, 1975).

Performing a backtest based on these conditional binomial probabilities should compare favourably to the original backtest proposed by Emmer et al., since the tested regions are actually independent.

Given that Tippett's method resembles Emmer's method most closely out of the combined p-value tests mentioned above, this would be a logical method to analyse, to see if any improvements over Emmer's performance can be found. The goal here is to find the appropriate mapping function such that extremely low values of S_T to correspond to low p-values of the Beta distribution (and thus rejection of false models), but once S_T goes up enough, we want the acceptance rate to increase fast enough as well (thus accepting true models consistently). Therefore, we can look at the shape of the Beta function for a variety of parameter combinations, compared to their acceptance rates in the setup that Wimmerstedt (2015) uses to analyse their performance. The distribution that Tippett suggests is: $S_T \sim Beta(1, m)$, with $m = 5$ in our case. This curve, along with several others, is shown in figure 14 on the left. We will not be very interested in what happens on the right side of this graph, since we will only be looking at whether or not the p-value exceeds the confidence level $\alpha = 95\%$. Thus, we zoom in on the rejection region on the right side of figure 14. We would like to analyse the Beta distributions that have the most change in curvature right around the point that they pass through the horizontal line, representing the rejection region, which corresponds to the following range of parameters: $a = 2;\ b \in \{4, ..., 8\}$.

However, the real problem of this method, and the original one by Emmer et al., is that we only use the information of the lowest of the $m = 5$ p-values, which technically incorporates some information on the others as well, but not nearly as much as we could be using. The issue that we must solve if we want to use all the information available (and thus find the most powerful test) is how we can combine

[6] A success is defined as any observation inside the x_i interval

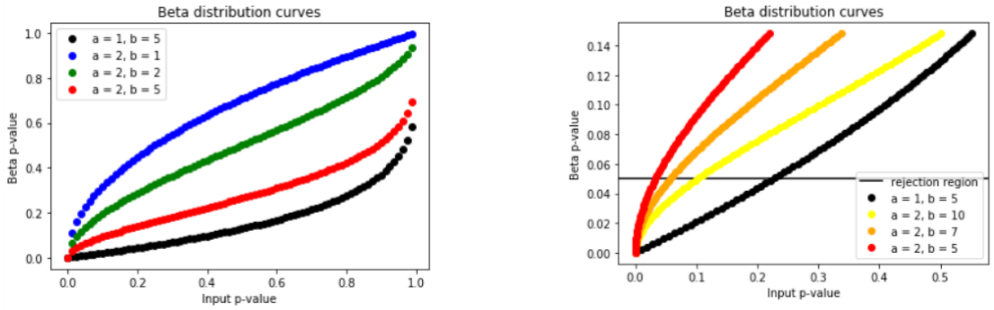

Figure 14: Beta distribution curves for several parameter combinations

these p-values in an appropriate and statistically correct manner. We want to be able to reject the null hypothesis even when only one of the intervals has an extreme enough amount of observations, or when several p-values are getting close to the α threshold, without actually exceeding it. What it comes down to, mathematically, is the issue of finding the appropriate mapping function of the information contained in a small sample of extreme events to a single p-value. Most of the previously mentioned methods, however, do not indicate rejection until at least several of the inputs are lower than our required confidence level α (which can be understood graphically from the paper by Heard and Rubin-Delanchy (2017)).

This leads me to consider tests that are rather different in their setup, perhaps relying on the symmetry and/or uniformity that we expect for our x_i exceedance levels.

When looking for a way to test the uniformity of the discrete distribution of our x_i values, there exists a number of test, such as the Chi-square goodness of fit test. For this test we can use the following formula:

$$\sum_{i=1}^{m} \frac{(x_i - E_i)^2}{E_i} \sim \chi^2_{m-1} \tag{54}$$

with m being equal to the amount of intervals to test, the 5 intervals as defined for x_i previously and the addition of x_6, the amount of observations in the interval $[VaR_{0.025}, \text{ inf}]$, thus $m = 6$. E_i represents the expected amount of observations in interval i, being equal to $0.005n$ for $i \in \{1, ..., 5\}$, and $0.98n$ for $i = 6$. Furthermore, the one-sided Chi-square test with $\theta = 5$ has a rejection threshold of $\chi^2_5 = 11.071$ at significance level $\alpha = 95\%$.

A.6.1 Emmer Extension results

We wish to know whether our proposed extension on the method by Emmer et al. (2015) outperforms the original. Therefore, we start our analysis with a simulation. We generate random draws (from the standard normal distribution) and test for rejection of the ES-estimate using the Emmer Extension. We do this using theoretical VaR-percentile thresholds that are correct for the standard normal distribution, given in table 13, thus testing the acceptance rate of true predictions. We also want to know the acceptance rates in a system where the true distribution is fat-tailed (but the assumed distribution is normal), thus the t_3 VaR-percentiles are also given.

VaR-percentile	z-value	t_3-value
2.5	1.96	3.18
2	2.06	3.48
1.5	2.17	3.90
1	2.33	4.54
0.5	2.58	5.84

Table 13: VaR-thresholds for given distributions

When a Monte Carlo simulation of 10^5 draws is run, we get results that should be directly comparable to those found in Wimmerstedt (2015); these are given in table 14. The methodology of Emmer et al. (2015) is also implemented, which gave an acceptance rate very close to the amount found by Wimmerstedt (which was 0.7793). We can see that the extension using Fisher's method clearly outperforms the original method in terms of accepting true scenarios. A sidenote here is that the acceptance rate is larger than our confidence level of 95%, which could be an indicator of overconfidence, since we would expect the random draw to be rejected in 5% of all simulation runs, by construction (using $\alpha = 5\%$). Another note here is that Emmer et al. (2015)'s method was notable for its underperformance in exactly this setting, compared to other backtests, and most other backtesting methods achieve acceptance rates very close to 95% with this setup (Wimmerstedt, 2015). Furthermore, acceptance rates are given for the method of Emmer et al. in the case that the 'proper' rejection threshold is used in a discrete setup, as explained in the literature section; results from using Tippett's method are given in table 15.

Method	Acceptance when true	Rejection when false
Emmer et al.	0.7808	0.9507
Emmer-Fisher	0.9890	0.2937
'Correct' Emmer	0.9095	0.8754
Emmer-Tippet	0.6938	0.9507

Table 14: Acceptance/rejection rates per backtesting model (for a random number of exceedances); for the false scenario $t_3(0, 1)$ is used

The analysis of rejections of false predictions will be discussed now. The first step in this setting is defining an incorrect prediction. As in Wimmerstedt (2015), we use draws from a t-distribution as the 'true' distribution, while still determining our ES-prediction using the standard normal distribution, as in the previous section. The results of this simulation are also shown in table 14. We see that Emmer et al.'s method performs very well here, but Fisher's method is lacking considerably in rejection power. A note can be made about the average p-value of Fisher's method, which equals 0.354. This shows that the combined method of Fisher does give doubt as to the validity of the ES-prediction, but it's nowhere near strict enough to actually reject the prediction consistently.

b	Acceptance when true	Rejection when false
4	0.9663	0.7491
5	0.9507	0.7491
6	0.9107	0.8728
7	0.8979	0.8728
8	0.8858	0.8728

Table 15: Acceptance/rejection rates using Tippett's method ($a = 1$)

We see that usage of the proper rejection threshold gives more conservative verdicts than the original in Wimmerstedt (2015), as expected, given that it has a significantly lower rejection rate in both the true scenario as well as the false scenario. Also, we can see that the use of Tippet's method does not do us much good, giving the same accuracy for rejection when false, but a lower acceptance rate when true.

The analysis of the goodness-of-fit tests will be considered next. Table 16 shows the true acceptance and false rejection rates just like tables 14 and 15 do.

Goodness-of-fit	Acceptance when true	Rejection when false
χ^2_5	0.9377	0.7490

Table 16: Acceptance/rejection rates using goodness-of-fit tests

Overall, this analysis shows that the extension using Fisher's method is better at recognising the correct ES-predictions than the original method by Emmer et al., but this is rather meaningless when it does not confidently reject ES-predictions that are clearly incorrect in a scenario of draws based on a t_3-distribution. Also, the other two methods show, at best, some improvement in one section, but worse rates in the other. Therefore, I suggest finding a different way of combining the 5 p-values from the thresholds, such that extreme events are weighed appropriately. Perhaps using symmetry and/or uniformity of the expected amounts x_1 through x_5, will give better rejection rates in false scenarios.

A.7 Generalised Pareto Distribution

The GPD is very useful for the modelling of tails of known continuous distributions. Throughout the literature on tail risk, this model is very prevalent. Therefore, we mention it here and it could prove useful for future analyses on the topic of tail risk management. Its CDF, given that the observation is an exceedance of the threshold $y < q_t(\nu)$ is given as follows:

$$F_t(y) = \begin{cases} F_{q_t(\nu),\xi,\beta_t}(y) = \nu\big(1 + \frac{\xi(q_t(\nu)-y)}{\beta_t}\big)^{-1/\xi}, & 0 < \xi < 1, \\ F_{q_t(\nu),\beta_t}(y) = \nu \, \exp\big(\frac{q_t(\nu)-y}{\beta_t}\big), & \xi = 0, \end{cases} \tag{55}$$

with $q_t(\nu)$ representing the VaR threshold at time t. Graham and Pál (2014) note that, in general, it is important to understand tail behaviour before using specific tail distribution models like the GPD. Thus, researchers are advised to perform due diligence to ensure that EVT concepts are appropriate for the specific setting, VaR confidence levels and portfolios used, before adopting these models.

References

Acerbi, C. & Szekely, B. (2014). Back-testing expected shortfall. *Risk*, *27*(11), 76–81.

Acerbi, C. & Szekely, B. (2017). General properties of a backtestable statistic.

Bayer, S. & Dimitriadis, T. (2018). Regression based expected shortfall backtesting. *arXiv preprint arXiv:1801.04112*.

BCBS. (2014). *A brief history of the basel committee.* Bank for International Settlements. Retrieved from `http://www.bis.org/bcbs/history.pdf`

BCBS. (2016). *Minimum capital requirements for market risk.* Bank for International Settlements. Retrieved from `https://www.bis.org/bcbs/publ/d352.htm`

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, *19*(4), 465–474.

Bradley, B. O. & Taqqu, M. S. (2003). Financial risk and heavy tails. *Handbook of Heavy-Tailed Distributions in Finance, ST Rachev, ed. Elsevier, Amsterdam*, 35–103.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 987–992.

Costanzino, N. & Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall.

Dalne, K. (2017). *The performance of market risk models for value at risk and expected shortfall backtesting: In the light of the fundamental review of the trading book.*

Del Brio, E. B., Mora-Valencia, A. & Perote, J. (2017). Risk quantification for commodity etfs: Backtesting value-at-risk and expected shortfall. *International Review of Financial Analysis*.

Dimitriadis, T. & Bayer, S. (2017). A joint quantile and expected shortfall regression framework. *arXiv preprint arXiv:1704.02213*.

Du, Z. & Escanciano, J. C. (2016). Backtesting expected shortfall: accounting for tail risk. *Management Science*, *63*(4), 940–958.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press.

Emmer, S., Kratz, M. & Tasche, D. (2015). What is the best risk measure in practice? a comparison of standard measures.

Fisher, R. A. (1934). Statistical methods for research workers. Edinburg: Oliver & Boyd.

Fissler, T., Ziegel, J. F. & Gneiting, T. (2015). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*.

Garcia-Jorcano, L. (2017). *Testing es estimation models: An extreme value theory approach* (Unpublished doctoral dissertation). Universidad Complutense de Madrid.

George, E. O. & Mudholkar, G. S. (1979). *The logit method for combining tests* (Tech. Rep.). ROCHESTER UNIV NY DEPT OF STATISTICS.

Graham, A. & Pál, J. (2014). Backtesting value-at-risk tail losses on a dynamic portfolio. *The Journal of Risk Model Validation*, *8*(2), 59.

Hall, P. (1927). The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, *19*(3/4), 240–245.

Heard, N. A. & Rubin-Delanchy, P. (2017). Choosing between methods of combining p-values.

Holton, G. A. (2014). *Value-at-risk: Theory and practice, second edition.* Retrieved from `https://www.value-at-risk.net`

Kerkhof, J. & Melenberg, B. (2004). Backtesting for risk-based regulatory capital. *Journal of Banking & Finance*, *28*(8), 1845–1865.

Lloyd, C. (2005, 12). Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation*, *75*, 921-934. doi: 10.1080/00949650412331321160

Löser, R., Wied, D. & Ziggel, D. (2018). New backtests for unconditional coverage of expected shortfall.

McNeil, A. J. & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, *7*(3-4), 271–300.

McNeil, A. J., Frey, R., Embrechts, P. et al. (2005). *Quantitative risk management: Concepts, techniques and tools* (Vol. 3). Princeton university press Princeton.

Moldenhauer, F. & Pitera, M. (2018). Backtesting expected shortfall: a simple recipe?

Nesmith, T. D., Oh, D. H. et al. (2017). Accurate evaluation of expected shortfall for linear portfolios with elliptically distributed risk factors. *Journal of Risk and Financial Management*, *10*(1), 5.

Nolde, N., Ziegel, J. F. et al. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The annals of applied statistics*, *11*(4), 1833–1874.

Righi, M. & Ceretta, P. (2013). Individual and flexible expected shortfall backtesting.

Righi, M. B. & Ceretta, P. S. (2015). A comparison of expected shortfall estimation models. *Journal of Economics and Business*, *78*, 14–47.

Sheikh, A. Z. & Qiao, H. (2009). Non-normality of market returns: A framework for asset allocation decision making. *The Journal of Alternative Investments*, *12*(3), 8–35.

Tippett, L. H. C. (1931). *Methods of statistics.* Williams Norgate: London.

Wimmerstedt, L. (2015). Backtesting expected shortfall: the design and implementation of different backtests.

Wong, W. K. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance*, *32*(7), 1404–1415.